特許協力条約に基づく国際出願願書 F005276W000

原本（出願用）- 印刷日時 2000年05月30日（30.05.2000）火曜日 14時16分19秒

0	受理官庁記入欄	
0-1	国際出願番号.	
0-2	国際出願日	
0-3	（受付印）	
0-4	様式-PCT/RO/101 この特許協力条約に基づく国際出願願書は、	
0-4-1	右記によって作成された。	PCT-EASY Version 2.90 (updated 08.03.2000)
0-5	申立て 出願人は、この国際出願が特許協力条約に従って処理されることを請求する。	
0-6	出願人によって指定された受理官庁	日本国特許庁（RO/JP）
0-7	出願人又は代理人の書類記号	F005276W000
I	発明の名称	文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体
II	出願人	
II-1	この欄に記載した者は	出願人である（applicant only）
II-2	右の指定国についての出願人である。	米国を除くすべての指定国（all designated States except US）
II-4ja	名称	セイコーエプソン株式会社
II-4en	Name	SEIKO EPSON CORPORATION
II-5ja	あて名:	163-0811 日本国 東京都 新宿区 西新宿２丁目４番１号
II-5en	Address:	4-1, Nishi-Shinjuku 2-Chome Shinjuku-Ku, Tokyo 163-0811 Japan
II-6	国籍（国名）	日本国 JP
II-7	住所（国名）	日本国 JP
II-8	電話番号	03-3348-3114
II-9	ファクシミリ番号	03-3340-4258
III-1	その他の出願人又は発明者	
III-1-1	この欄に記載した者は	出願人及び発明者である（applicant and inventor）
III-1-2	右の指定国についての出願人である。	米国のみ（US only）
III-1-4ja	氏名(姓名)	長石 道博
III-1-4en	Name (LAST, First)	NAGAISHI, Michihiro
III-1-5ja	あて名:	392-8502 日本国 長野県 諏訪市 大和３丁目３番５号 セイコーエプソン株式会社内
III-1-5en	Address:	c/o SEIKO EPSON CORPORATION 3-5, Owa 3-Chome, Suwa-Shi, Nagano 392-8502 Japan
III-1-6	国籍（国名）	日本国 JP
III-1-7	住所（国名）	日本国 JP

特許協力条約に基づく国際出願願書 F005276W000
原本（出願用）- 印刷日時 2000年05月30日（30.05.2000）火曜日 14時16分19秒

III-2	その他の出願人又は発明者	
III-2-1	この欄に記載した者は	出願人及び発明者である（applicant and inventor）
III-2-2	右の指定国についての出願人である。	米国のみ（US only）
III-2-4ja	氏名(姓名)	三輪 真司
III-2-4en	Name (LAST, First)	MIWA, Shinji
III-2-5ja	あて名:	392-8502 日本国 長野県 諏訪市 大和３丁目３番５号 セイコーエプソン株式会社内
III-2-5en	Address:	c/o SEIKO EPSON CORPORATION 3-5,Owa 3-Chome, Suwa-shi, Nagano 392-8502 Japan
III-2-6	国籍（国名）	日本国 JP
III-2-7	住所（国名）	日本国 JP
IV-1	代理人又は共通の代表者、通知のあて名 下記の者は国際機関において右記のごとく出願人のために行動する。	代理人（agent）
IV-1-1ja	氏名(姓名)	鈴木 喜三郎
IV-1-1en	Name (LAST, First)	SUZUKI, Kisaburo
IV-1-2ja	あて名:	392-8502 日本国 長野県 諏訪市 大和３丁目３番５号 セイコーエプソン株式会社　知的財産部内
IV-1-2en	Address:	c/o Intellectual Property Department SEIKO EPSON CORPORATION 3-5, Owa 3-chome Suwa-Shi, Nagano 392-8502 Japan
IV-1-3	電話番号	0266-52-3139
IV-1-4	ファクシミリ番号	0266-58-3243
IV-2	その他の代理人	筆頭代理人と同じあて名を有する代理人（additional agent(s) with same address as first named agent）
IV-2-1ja	氏名	上柳 雅誉; 須澤 修
IV-2-1en	Name(s)	KAMIYANAGI, Masataka; SUZAWA, Osamu
V	国の指定	
V-1	広域特許 (他の種類の保護又は取扱いを求める場合には括弧内に記載する。)	AP: GH GM KE LS MW SD SL SZ TZ UG ZW 及びハラレプロトコルと特許協力条約の締約国である他の国 EA: AM AZ BY KG KZ MD RU TJ TM 及びユーラシア特許条約と特許協力条約の締約国である他の国 EP: AT BE CH&LI CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE 及びヨーロッパ特許条約と特許協力条約の締約国である他の国 OA: BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG 及びアフリカ知的所有権機構と特許協力条約の締約国である他の国

特許協力条約に基づく国際出願願書 F005276W000

原本（出願用）- 印刷日時 2000年05月30日（30.05.2000）火曜日 14時16分19秒

V-2	国内特許 (他の種類の保護又は取扱いを求める場合には括弧内に記載する。)	AE AG AL AM AT AU AZ BA BB BG BR BY CA CH&LI CN CR CU CZ DE DK DM DZ EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ UA UG US UZ VN YU ZA ZW	
V-5	指定の確認の宣言 出願人は、上記の指定に加えて、規則4.9(b)の規定に基づき、特許協力条約のもとで認められる他の全ての国の指定を行う。ただし、V-6欄に示した国の指定を除く。出願人は、これらの追加される指定が確認を条件としていること、並びに優先日から15月が経過する前にその確認がなされない指定は、この期間の経過時に、出願人によって取り下げられたものとみなされることを宣言する。		
V-6	指定の確認から除かれる国	なし（NONE）	
VI-1	先の国内出願に基づく優先権主張		
VI-1-1	先の出願日	1999年06月04日（04.06.1999）	
VI-1-2	先の出願番号	特願平11-158498	
VI-1-3	国名	日本国 JP	
VI-2	先の国内出願に基づく優先権主張		
VI-2-1	先の出願日	1999年07月27日（27.07.1999）	
VI-2-2	先の出願番号	特願平11-212501	
VI-2-3	国名	日本国 JP	
VI-3	優先権証明書送付の請求 上記の先の出願のうち、右記の番号のものについては、出願書類の認証謄本を作成し国際事務局へ送付することを、受理官庁に対して請求している。	VI-1, VI-2	
VII-1	特定された国際調査機関(ISA)	日本国特許庁（ISA/JP）	
VIII	照合欄	用紙の枚数	添付された電子データ
VIII-1	願書	4	-
VIII-2	明細書	37	-
VIII-3	請求の範囲	4	-
VIII-4	要約	1	f005276wo00.txt
VIII-5	図面	14	-
VIII-7	合計	60	
	添付書類	添付	添付された電子データ
VIII-8	手数料計算用紙	✓	-
VIII-9	別個の記名押印された委任状	✓	-
VIII-16	PCT-EASYディスク	-	フレキシブルディスク
VIII-17	その他	納付する手数料に相当する特許印紙を貼付した書面	-
VIII-18	要約書とともに提示する図の番号	1	
VIII-19	国際出願の使用言語名:	日本語（Japanese）	

特許協力条約に基づく国際出願願書 F005276W000

原本（出願用）- 印刷日時 2000年05月30日（30.05.2000）火曜日 14時16分19秒

IX-1	提出者の記名押印	
IX-1-1	氏名(姓名)	鈴木 喜三郎
IX-2	提出者の記名押印	
IX-2-1	氏名(姓名)	上柳 雅誉
IX-3	提出者の記名押印	
IX-3-1	氏名(姓名)	須澤 修

受理官庁記入欄

10-1	国際出願として提出された書類の実際の受理の日	
10-2 10-2-1 10-2-2	図面 : 受理された 不足図面がある	
10-3	国際出願として提出された書類を補完する書類又は図面であってその後期間内に提出されたものの実際の受理の日（訂正日）	
10-4	特許協力条約第11条(2)に基づく必要な補完の期間内の受理の日	
10-5	出願人により特定された国際調査機関	ISA/JP
10-6	調査手数料未払いにつき、国際調査機関に調査用写しを送付していない	

国際事務局記入欄

11-1	記録原本の受理の日	

明細書

文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体

背景技術

本発明は多数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法および文書分類装置並びに文書分類処理プログラムを記録した記録媒体に関する。

多数の文書を意味的なまとまりごとの複数のクラスタに分類する際、それぞれの文書から特徴要素を抽出し、その特徴要素に基づいて分類することが行われている。その分類手法として、それぞれの文書全体（表題や本文など1つの文書を構成する文書内容全体）を特徴要素の抽出対象とし、それぞれの文書全体から特徴要素を抽出し、抽出された特徴要素に基づいて複数のクラスタに分類する文書分類方法がある。

この文書全体を特徴要素抽出の対象として分類を行うと、文書の形態素解析や、特徴抽出処理が非常に繁雑であり、情報処理装置において、中央処理装置（ＣＰＵ）がその処理を行う場合、ＣＰＵに対する負荷を大きいものとしている。また、一般に、文書はその文書の主旨とは直接関係のない記述を多く含んでいるのが普通である。したがって、文書全体を特徴要素抽出の対象とすると、それによって分類されたクラスタは情報の分類という観点から見たとき、あまり意味のない分類となることも多い。つまり、ノイズクラスタが多数生成されてしまうということにもなる。

このような問題点を解消する手法として、それぞれの文書の主旨を適切に表す部分としてそれぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法がある。

この手法は、文書の主旨を適切に反映した文書分類を可能とすることができるものとして期待されている。

このように、文書を幾つかのクラスタに分類する手法は幾つか考えられている。

しかしながら、上述した文書の表題部から抽出された特徴要素に基づいて文書を分類する手法を用いたとしても、それによって得られる分類結果は、クラスタの数が多くなりすぎることもあり、ユーザ側から見たときに、決して適切な分類が行われたとは思えない場合もでてくる。例えば、分類されて得られる多数のクラスタを比較した場合、それぞれのクラスタに共通した文書が数多く含まれる場合もある。このような場合、ユーザは提示された多数のクラスタについて、結局は、自分で整理し、その中から自分の本当に欲しい情報を探すというような面倒な処理を行うことになる。

そこで、本発明は、分類結果として得られた多数のクラスタに対してクラスタマージ処理を行うことで、より一層、ユーザにとってわかりやすく簡潔的に分類された分類結果を提示できるようにすることを目的としている。

図面の簡単な説明

図１は、本発明の第１の実施形態を説明するブロック図である。

図２は、本発明の第１の実施形態を説明するための複数の文書例を示す図である。

図３は、本発明の第１の実施形態における文書分類処理の処理手順を概略的に説明するフローチャートである。

図４は、特徴要素とそれぞれの文書との関係を示す特徴テーブル内容の一例を示す図である。

図５は、図４に示す特徴テーブルに基づいて文書を分類した分類結果を示す図である。

図６は、２つのクラスタ間でのクラスタマージ処理を説明する図であり、それぞれのクラスタに含まれる文書例を示す図である。

図７は、図５の分類結果についてクラスタマージ処理した結果を示す図である。

図８は、特徴要素が元の文書にどのように出現するかによってクラスタマージを行う場合の文書分類装置のブロック図である。

図９は、本発明の第２の実施形態を説明するブロック図である。

図１０は、本発明の第２の実施形態を説明するための複数の文書例を示す図で

ある。

図１１は、本発明の第２の実施形態において行う文書分類処理の処理手順を概略的に説明するフローチャートである。

図１２は、特徴要素とそれぞれの文書との関係を示す特徴テーブル内容の一例を示す図である。

図１３は、図１２に示す特徴テーブルに基づいて文書を分類した分類結果を示す図である。

図１４は、２つのクラスタ間でのクラスタマージ処理を説明する図であり、それぞれのクラスタに含まれる文書例を示す図である。

図１５は、図１３の分類結果についてクラスタマージ処理した結果を示す図である。

図１６は、クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタのクラスタ名をＡＮＤ形式（横一列に並べた場合）の表記の仕方で表示した例を示す分類結果例を示す図である。

図１７は、クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタのクラスタ名をＡＮＤ形式（クラスタ名を１つづつ縦に並べた場合）の表記の仕方で表示した例を示す分類結果例を示す図である。

発明の開示

前述の目的を達成するために、本発明の文書分類方法は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行うようにしている。

また、前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージする処理である。

また、前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラ

スタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージする処理であってもよい。

そして、これらクラスタマージ処理は、少なくとも２つのクラスタ間で行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すようにする。

さらに、前記クラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力する。

このように本発明は、それぞれの文書を複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書の内容に基づいてそれぞれのクラスタ間の関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行うようにしている。これによって、最初のクラスタリング処理によって、多数のクラスタが生成されたとしても、それぞれのクラスタ間でクラスタ同志の関連性を判断し、関連性の高い複数のクラスタを統合することができるので、簡潔化された分類結果をユーザに提示することができ、ユーザは自分の欲しい情報を効率よく探すことができるようになる。

また、クラスタ間の関連性の判断は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、各々のクラスタに共通して含まれる文書数を基にして行うので、簡単で的確なクラスタマージ処理を行うことができる。

また、クラスタ間の関連性の判断を行うための他の方法として、特徴要素がクラスタマージ処理対象となるクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージ処理を行うようにしてもよく、これによれば、実際の文書内容に基づいたクラスタ同志の関連性の判断が行えるので、適切なクラスタマージ結果を得ることができる。

そして、クラスタマージ処理は、少なくとも２つのクラスタの組み合わせで行い、さらに、所定の数のクラスタ間でのクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことによって、最終的には、

より簡潔的に整理された分類結果を得ることができる。

また、このようなクラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することにより、ユーザはどのような状況でクラスタマージ処理がなされたかを知ることができるので、クラスタマージ処理後の結果から自分の欲しい情報を探す際に、その付加状況を参考にして探すことができる。

本発明の第２の文書分類方法は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行い、このクラスタマージ処理によって得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、その表示内容をユーザに提示すべき分類結果に含めて出力するようにしている。

また、前記クラスタマージ処理内容がわかるような表示内容とは、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示内容であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をＡＮＤ形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をＯＲ形式の表記の仕方で表示させるようにしている。

そして、前記ＡＮＤ形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記ＯＲ形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記するようにしている。

さらに、あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記

の仕方で表示することも可能としている。

このように本発明は、クラスタマージされて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、それを表示するようにしている。

これによって、ユーザは、クラスタマージされる前のクラスタの様子、すなわち、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかといったことを表示内容を見るだけで知ることができる。そして、どのような関連性を有しているかを示す表示の仕方としては、クラスタマージ処理されて得られた新たなクラスタに含まれるクラスタ間の関連性の高さに基づき、クラスタマージ処理されたそれぞれのクラスタのクラスタ名の表示の仕方を変えるようにしている。

そのクラスタ名の表示の仕方は、具体的には、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をＡＮＤ形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さを表す値が予め設定された値未満である場合には、それぞれのクラスタ名をＯＲ形式の表記の仕方で表示させるようにしている。たとえば、関連性の高さがきわめて高い場合には、それぞれのクラスタ名を横一列に連続的に表示したり、それぞれのクラスタ名を１つづつ縦に並べて表示し、関連性の高さがそれほどでもない場合には、それぞれのクラスタ名の間に区切り記号を挿入するなどして表示する。ユーザはこのような表示を見ることで、統合される前のそれぞれのクラスタがどのようなクラスタであって、それぞれのクラスタ同志の関連性がどの程度であるかなどを知ることができる。

また、あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することも可能であり、包含関係であることを繁雑なイメージを使わないでもわかりやすく表示できる。

また、本発明の文書分類装置は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、前記複数の文書を意味的に共通性を

有する複数のクラスタに分類するクラスタリング部と、このクラスタリング部により得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部とを有する構成としている。

また、本発明の文書分類装置は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、このクラスタリング部によって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ部と、このクラスタマージ部によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ内容生成部と、その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力手段とを有した構成としている。

また、本発明の文書分類処理プログラムを記録した記録媒体は、複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類処理プログラムを記録した記録媒体であって、その文書分類処理プログラムは、前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング処理手順と、これにより分類された複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理手順とを含むものである。

さらに、本発明の文書分類処理プログラムを記録した記録媒体は、複数の文書を意味的に共通性を有する複数のクラスタに分類して出力する文書分類処理プログラムを記録した記録媒体であって、その処理プログラムは、複数の文書を意味的に共通性を有する複数のクラスタに分類する手順と、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも2つのクラスタを統合するクラスタマージ処理を行う手順と、クラスタマージ処理されて得られた新たなクラスタの

表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成する手順と、その表示内容をユーザに提示すべき分類結果に含めて出力する手順とを含むようにしている。

発明を実施するための最良の形態

（第１の実施形態）

以下、本発明の第１の実施形態について説明する。なお、この実施形態で説明する内容は、本発明の文書分類方法および文書分類装置についての説明であるとともに、本発明の文書分類処理プログラムを記録した記録媒体における文書分類処理プログラムの具体的な処理内容をも含むものである。

また、この実施形態では、文書分類の手法として、前述したように、それぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法を用いるものとする。

図１は本実施形態の装置構成を示すもので、大きく分けると、複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部１と、このクラスタリング部１により得られた複数のクラスタ間で各々のクラスタに含まれる文書の内容に基づいて各々のクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ部２と、このクラスタマージ部２でクラスタマージ処理された分類結果を出力する分類結果出力部３とを有した構成となっている。

クラスタリング部１は、文書記憶部１１、文解析部１２、特徴要素抽出部１３、特徴テーブル作成部１４、文書分類部１５、分類結果記憶部１６を有している。

クラスタマージ部２はクラスタを統合するものであるがこれについての処理内容については後に詳細に説明する。

分類結果出力部３は、出力制御部３１、表示部３２を有し、クラスタマージ部２によるクラスタマージ処理結果を出力させるための制御を行う。

上述のクラスタリング部１に含まれる文書記憶部１１はこの場合、多数の文書データをデータベースとして持つものである。ここでは、たとえば、図２に示す

ような文書群を分類する場合を説明する。図２に示される文書群は、それぞれが独立した文書Ｄ１，Ｄ２，・・・，Ｄ７を有し、これらの文書Ｄ１，Ｄ２，・・・，Ｄ７は表題部Ｔ１，Ｔ２，・・・，Ｔ７と、それに対する本文Ａ１，Ａ２，・・・，Ａ７を持っているものとする。

文解析部１２は文書記憶部１１に記憶されている文書を文解析し、それぞれの文書の表題部を検出する。この文解析部１２が行う表題部の検出は、具体的には次のようにして行う。

まず、第１の方法として、文書構造様式によって表題と規定される部分があればその部分を表題部とする。また、第２の方法として、文書構造様式によって、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とする。また、第３の方法として、定められた数の文または単語を文書先頭より抽出し、その抽出した部分を表題部とする。さらには、これら第１、第２、第３の方法を順次行い、第１の方法を行ったとき、表題と規定されている部分があればその部分を表題部とし、表題と規定される部分が存在しなければ、第２の方法を行い、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とし、標準より大きな文字で表示する指定がなされていなければ、第３の方法を行って表題部を検出する。

特徴要素抽出部１３は、文解析部２で検出されたそれぞれの文書の表題部の中から特徴要素を抽出する。

特徴テーブル作成手段１４は、前記表題部から抽出された特徴要素とそれぞれの文書との関係を示す特徴テーブルを作成する。なお、この特徴テーブルの具体的な内容については後述する。

文書分類部１５は、前述の特徴テーブルの内容を参照し、文書Ｄ１，Ｄ２，・・・，Ｄ７を意味的に共通性のある複数のクラスタに分類する。つまり、文書Ｄ１，Ｄ２，・・・，Ｄ７の表題部に存在する特徴要素に基づいて、共通する特徴要素を持つ処理対象文書を１つのまとまりとし、そのまとまりを１つのクラスタとする。なお、この文書分類部１５は同義特徴辞書（図示せず）を有し、共通する特徴要素を持つ処理対象文書を１つのまとまりとする処理を行う際、共通する特徴要素であるか否かの判断を、その同義語辞書を用い同義語が有るか否かにより行い、

同義語が存在する場合にはそれを同じクラスタとする処理を行うことも可能である。

分類結果記憶部１６は、文書分類部１５によって分類された内容を記憶する。

このような構成において、本発明の文書分類処理について説明する。本実施形態においては、文書分類処理は、図３のフローチャートに示すように、処理対象となる多数の文書を意味的に共通性を有する複数のクラスタに分類し（ステップＳ１）、これにより分類された複数のクラスタ間で各々のクラスタに含まれる文書に基づいて（これについては後に説明する）それぞれのクラスタの関連性を判断する（ステップＳ２）。そして、一定以上の関連性を有する少なくとも２つのクラスタを統合する（ステップＳ３）。以下、具体例を参照して詳細に説明する。

ここでは、図２で示した文書Ｄ１，Ｄ２，・・・，Ｄ７を分類する例について説明する。この実施の形態では、それぞれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理された結果についてクラスタマージ処理を行う。まず始めに、表題部から特徴要素を抽出し、その抽出された特徴要素に基づいて行われるクラスタリング処理（クラスタリング部１が行う処理）について説明する。

これらの文書Ｄ１，Ｄ２，・・・，Ｄ７は、文解析部１２にて表題部が検出される。たとえば、文書Ｄ１については表題部Ｔ１が検出され、文書Ｄ２については表題部Ｔ２が検出され、文書Ｄ３については表題部Ｔ３が検出されるというように、それぞれの文書Ｄ１，Ｄ２，・・・，Ｄ７の表題部Ｔ１，Ｔ２，・・・，Ｔ７が検出される。

そして、特徴要素抽出部１３によって、それぞれの表題部に存在する特徴要素が抽出されたのち、特徴テーブル作成部１４により、それぞれの特徴要素とその特徴要素を表題部に含む文書との関係を示す特徴テーブルが作成される。この特徴テーブルの例を図４に示す。なお、ここでは、文書数が３つ以上取り出される特徴要素とその特徴要素を含む文書との関係を示し、特徴テーブル内に示される数値は、その特徴要素が各文書の表題部に幾つ含まれるているかの数を示している。たとえば、「用紙」という特徴要素は、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７のそれぞれの表題部に、それぞれ１個ずつ含まれていることを示している。

図４の特徴テーブルからもわかるように、表題部に「用紙」という特徴要素を含む文書は、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７であり、また、表題部に「カセット」という特徴要素を含む文書は、文書Ｄ１，Ｄ４，Ｄ７であり、さらに、表題部に「増設」という特徴要素を含む文書は、文書Ｄ２，Ｄ３，Ｄ５，Ｄ７である。なお、図２において、これら各特徴要素部分にはアンダーラインが施されている。

そして、文書分類部１５はこのような特徴テーブルを参照して、それぞれの特徴要素ごとの文書クラスタ分けを行う。その分類結果を図５に示す。なお、このようなクラスタに分類する際、前述したように、共通する特徴要素であるか否かの判断を、同義語辞書を用い同義語が有るか否かによっても行い、同義語が存在する場合にはそれを同じ文書クラスタとする処理を行うことも可能である。たとえば、「用紙」と「印刷紙」の両方が特徴要素として抽出されたとすれば、これらの特徴要素を表題部に含む文書は同じクラスタとするなどという処理を行う。

このような分類結果は分類結果記憶部１６に格納される。図５に示される分類結果において、たとえば、「用紙」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれる）について見れば、図２の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ６は印刷された後の用紙の汚れについての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７はどれも用紙に関する内容であり、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

また、「カセット」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ７が含まれる）について見れば、図２の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７にはどれも用紙をセットすることに関する内容が含まれており、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

また、「増設」で分類されたクラスタ（文書Ｄ２，Ｄ３，Ｄ５，Ｄ７が含まれ

る）について見れば、図２の文書内容からもわかるように、文書Ｄ２はメモリの増設についての内容であり、文書Ｄ３はインタフェースカードの増設についての内容であり、文書Ｄ５はハードディスクの増設についての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ２，Ｄ３，Ｄ５，Ｄ７はどれも何かを増設する場合についての内容であり、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

このような適切な分類が行える理由としては、それぞれの文書の表題部から特徴要素を抽出し、その特徴要素に基づいて文書を分類しているからである。つまり、文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることが多い。したがって、文書の表題部に含まれる特徴要素を用いて分類を行うことにより、分類結果が散漫になることが少なく、また、ノイズクラスタが生成される率も少なくすることができる。また、各文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることから、文書の制作者側の視点による分類が得られる。

そして、分類が行われた後、ユーザによって、たとえば、「用紙」についてのクラスタの選択指示が出されたとすると、そのクラスタに属する文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が文書記憶部１１から読み出されて表示部３２に表示される。なお、このときの表示内容としては、前述したように、文書番号や文書名のみでもよく、さらには、その文書内容を表示させるようにしてもよい。

ところで、本発明は以上のようにクラスタリング処理した結果について、さらに、クラスタマージ部２によってクラスタマージ処理を行う。

すなわち、図５に示す分類結果において、特徴要素である「用紙」と「カセット」について見ると、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれ、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７に存在することがわかる。

このように、「用紙」のクラスタと「カセット」のクラスタには、共に文書Ｄ１，Ｄ４，Ｄ７が共通して存在している。これは、「用紙」という特徴要素と「カセット」という特徴要素は相互に関連した状態で用いられることが多いことを意

味している。たとえば、文書Ｄ１，Ｄ４，Ｄ７の表題部または本文のなかに「用紙カセット」という用語が用いられている。つまり、これらの文書Ｄ１，Ｄ４，Ｄ７は共通性の高い文書であり、これら文書Ｄ１，Ｄ４，Ｄ７は同じクラスタに分類した方がより好ましいと考えられる。

これを実現するために本発明では特徴要素に基づいてクラスタリングしたあと、そのクラスタリング結果に対しクラスタマージ処理を施す。

このクラスタマージ処理について以下に説明する。まず始めに、図５の分類結果とは関係なく一般的な例について図６を参照しながら説明する。

今、２つのクラスタＣ１，Ｃ２があるとする。クラスタＣ１として５個の文書Ｄ１，Ｄ２，Ｄ３，Ｄ４，Ｄ８が抽出され、クラスタＣ２には６個の文書Ｄ３，Ｄ４，Ｄ５，Ｄ６，Ｄ７，Ｄ８が抽出されたとする。

ここで、２つのクラスタＣ１，Ｃ２に共通している文書は、文書Ｄ３，Ｄ４，Ｄ８である。この実施の形態では、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基に、それぞれのクラスタ間の関連性を判断してクラスタマージ処理を行う。

具体的には、複数のクラスタととして、ある２つのクラスタに共通している文書数が２つのクラスタに存在する合計の文書数に対しどのくらいの割合かを計算し、その計算結果が予め定めたしきい値以上かどうかによってマージするか否かを決める。

たとえば、この場合、２つのクラスタＣ１，Ｃ２に存在する文書数の合計は１１個であり、両者に共通する文書数は３個である。これらから合計の文書数に占める共通する文書数の割合（％）を計算し、その結果からマージするか否かを決定する。この割合（％）を求める際、合計の文書数で共通する文書数を単純に割り算してそれに１００を掛けて求めてもよいが、共通する文書数に任意に設定される係数を掛け算したものを合計の文書数で割り算してそれに１００を掛けて求めるようにしてもよい。

一例として、クラスタＣ１に存在する文書数を$\alpha 1$、クラスタＣ２に存在する文書数を$\alpha 2$とし、両者に共通する文書数をβとした場合、たとえばβに係数としてたとえば２を掛けて、$2\beta/(\alpha 1+\alpha 2)\times 100$を計算し、その値（％）

が予め設定されたしきい値ＴＨ（％）と比較して、上式による計算結果がしきい値ＴＨ以上であればマージするというようなことを行う。図６で示した例について考えれば、２βは２×３＝６個、α１＋α２は５＋６＝１１個であるので、この場合、約５５％と求められる。ここで、しきい値ＴＨが仮に７０％と設定されているとすれば、計算結果（５５％）はしきい値ＴＨ（７０％）より小さいので、クラスタＣ１とクラスタＣ２はマージしないとする。なお、係数は任意に設定されるもので、計算結果で得られる数値（％）がしきい値と比較し易いような値となるように適当に設定されるものであり、この場合は係数を２としたが、係数を１としても特に問題はない。

ここで、図５で示した分類結果を例にして説明すれば、図５の場合、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７の３つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ１，Ｄ４，Ｄ７の３つの文書であり、これを合計の文書数に対する割合（％）で考える。

これを前述した計算式によって計算する。図５の分類結果の場合、合計の文書数（α１＋α２）は、４＋３＝７となり、共通の文書数は３で２βは６となる。したがって、この場合、約８６％という高い値が得られる。これは、設定されたしきい値（ここでは７０％としている）よりも高いので、この「用紙」のクラスタと「カセット」のクラスタはマージして１つのクラスタとするということになる。

同様に考えて、図５の「用紙」のクラスタと「増設」のクラスタとをマージするか否か、「カセット」のクラスタと「増設」のクラスタとをマージするか否かについて判断する。

まず、「用紙」のクラスタと「増設」のクラスタについては、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを上式を用いて計算すると、この場合、２５％という結果が得られ、これは、しきい値（７０％）よりも低い値であるので、この場合は、両者はマージしないとする。

また、「カセット」のクラスタと「増設」のクラスタについては、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７の３つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを上式を用いて計算すると、この場合、約２８％という結果が得られ、これは、しきい値（７０％）よりも低い値であるので、この場合は、両者はマージしないとする。

このようにして、それぞれのクラスタに対し２つのクラスタごとにそれぞれマージするか否かを判断する。この図５の分類結果についてマージするか否かの処理を行ったあとの分類結果（マージ処理後の分類結果という）が図７である。図７によれば、「用紙」と「カセット」が「用紙＋カセット」という１つのクラスタに分類され、そのクラスタに属する文書は文書Ｄ１，Ｄ４，Ｄ６，Ｄ７ということになる。また、「増設」についてはそのまま単独のクラスタを構成する。

図７に示されるクラスタマージ処理後の分類結果において、たとえば、「用紙＋カセット」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれる）について見れば、図２の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ６は印刷された後の用紙の汚れた場合にはどのようにするかについての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７はどれも用紙やカセットに関する内容であり、１つのクラスタとして分類されて何等問題のないものとなり、むしろ、「用紙＋カセット」を１つのクラスタとした方がよい分類結果であるといえる。

このように、始めにそれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理されて得られたそれぞれのクラスタに対し、２つづつのクラスタの組み合わせについてクラスタマージ処理を行うことによって、より適切なクラスタリングが行える。

また、以上のようにして２つのクラスタごとに１回目のクラスタマージ処理が終了し、図７のようなクラスタマージ処理後の分類結果が得られると、今度は、

そのクラスタマージ処理後の分類結果について、２回目のクラスタマージ処理を行う。つまり、図７の１回目のクラスタマージ処理後の結果で考えた場合、「用紙＋カセット」のクラスタと「増設」のクラスタについてクラスタマージ処理を行う。この場合、「用紙＋カセット」のクラスタと「増設」のクラスタについては、「用紙＋カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを合計の文書数に対する割合（％）で考えると、共通する文書数１に定数２を掛けたものを合計の文書数８で割り算し、それに１００を掛けると、２５％という結果が得られ、これは、しきい値（７０％）よりも低い値であるので、この場合は、両者はマージしないとする。

このようにして、２つのクラスタ間で１回目のクラスタマージ処理が終了した後、その１回目のクラスタマージ処理に新たな２つのクラスタ間で２回目のクラスタマージ処理を行い、その２回目のクラスタマージ処理が終了した後、その２回目のクラスタマージ処理後に新たな２つのクラスタ間で３回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなるまで）その処理を繰り返す。

また、これまでの説明は、２つのクラスタ間でクラスタマージ処理を行う例についてであるが、クラスタマージ処理は３つ以上のクラスタの組み合わせについても可能である。この場合、１回のクラスタマージ処理によって３つ以上のクラスタ間でクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である。なお、３つ以上のクラスタについてクラスタマージするか否かを判断する場合、前述したように、それぞれのクラスタに存在する合計の文書数に対する共通の文書数の割合（％）で考えることができる。

さらに、これまで説明した複数のクラスタ間でのクラスタマージ処理は、図５に示すような分類結果に基づき、それぞれのクラスタ間に共通する文書数が合計の文書数に占める割合を求め、それを設定されたしきい値との比較によって求めるようにしたが、このような方法によらず、それぞれのクラスタを特徴づける特

徴要素が、元の文書においてどのような状態で用いられているかを調べることによってもクラスタマージ処理を行うことができる。これを実現するための文書分類装置の構成例を図８に示す。図８に示されるそれぞれの構成要素は図１と同じであり、同一部分には同一符号が付されているが、この場合、元の文書内容からクラスタマージするか否かを判断するため、クラスタマージ部２には、文書記憶部１１の出力が与えられるようになっている。以下、これについて説明する。

図５に示すような分類結果において、「用紙」のクラスタと「カセット」のクラスタをクラスタマージ処理する場合について説明する。「用紙」のクラスタには、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれ、「カセット」のクラスタには、文書Ｄ１，Ｄ４，Ｄ７が含まれる。

これら文書おいて、「用紙」と「カセット」がどのように用いられているかを調べる。文書Ｄ１においては、「用紙」と「カセット」が結びついた「用紙カセット」という用語が複数箇所出現し、文書Ｄ４には文書Ｄ１と同様に「用紙カセット」という用語が存在するとともに、「用紙」と「カセット」が近接した状態で用いられている。また、文書Ｄ７にも「用紙カセット」という用語や「用紙カセットユニット」という用語が存在する。また、文書Ｄ６には「カセット」という用語は存在しないが「用紙」という用語が複数出現する。

これらのことから考えれば、特徴要素として抽出された「用紙」と「カセット」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素であり、両者は関連性の高い特徴要素であることがわかる。このことから、少なくとも文書Ｄ１，Ｄ４，Ｄ７は関連性の高い文書であり、文書Ｄ６も全く関連性がないとは言えないので、この場合、「用紙」のクラスタと「カセット」のクラスタは「用紙＋カセット」のクラスタとして１つにまとめても問題がないと判断できる。

次に、「用紙」のクラスタと「増設」のクラスタをクラスタマージ処理する。「用紙」のクラスタには、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれ、「増設」のクラスタには、文書Ｄ２，Ｄ３，Ｄ５，Ｄ７が含まれる。

これら文書おいて、「用紙」と「カセット」がどのように用いられているかを調べる。文書Ｄ１、Ｄ２，Ｄ３，Ｄ４，Ｄ５，Ｄ６においては、「用紙」と「増設」が結びついて用いられた部分や、近接して用いられている部分はなく、文書

Ｄ７のみにおいて「用紙カセット」と「増設」が近接した状態で用いられている程度である。

したがって、これらのことから、特徴要素として抽出された「用紙」と「増設」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素ではなく、両者はあまり関連性のある特徴要素であるとはいえないことがわかる。このことから、「用紙」のクラスタと「増設」のクラスタはマージしない方がよいということがわかる。

また、「カセット」のクラスタと「増設」のクラスタをクラスタマージ処理すると、この場合も、「用紙」のクラスタと「増設」のクラスタにおけるクラスタマージ処理と同様に、「カセット」と「増設」が結びついて用いられた部分や、近接して用いられている部分は少ない。

したがって、これらのことから、特徴要素として抽出された「カセット」と「増設」は、連続的に用いられたり近接して用いられたりすることの多い特徴要素ではなく、両者はあまり関連性のある特徴要素であるとはいえないことがわかる。このことから、「カセット」のクラスタと「増設」のクラスタはマージしない方がよいということがわかる。

なお、このようなそれぞれのクラスタを特徴づける特徴要素が元の文書においてどのような状態で存在するかによってクラスタマージする処理においても、前述したように、それぞれのクラスタ間で１回目のクラスタマージ処理が終了した後、その１回目のクラスタマージ処理後に新たなクラスタ間で２回目のクラスタマージ処理を行い、その２回目のクラスタマージ処理が終了した後、その２回目のクラスタマージ処理後に新たなクラスタ間で３回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなるまで）その処理を繰り返す。

また、この場合も２つのクラスタ間でのクラスタマージ処理だけでなく、クラスタマージ処理は３つ以上のクラスタの組み合わせについても可能である。この場合、１回のクラスタマージ処理によって３つ以上のクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である。

ところで、以上のようにしてクラスタマージ処理を行ったあと、クラスタマージされた後の結果をユーザに表示する際、どのような状況でクラスタマージを行ったのかを示す情報を付加情報としてユーザに提示することが好ましい。これは、クラスタマージ部２で行った処理内容を出力制御部３１が受けてそれを表示部３２に表示させるようにすることで行える。

なお、本実施形態は、上記内容に限定されるものではなく、上記の要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、前述の実施の形態では、図５に示すような分類結果を得るための特徴要素を各文書の表題部から得るようにして、表題部から得られた特徴要素に基づいたクラスタリングを行う例について説明したが、本実施形態においては、複数の文書を意味的に共通性のあるクラスタに分類し、その分類結果についてクラスタマージ処理を行うものであるので、複数の文書をクラスタリングする手法は、特に限定されるものではない。複数の文書をクラスタリングする手法としては、前述の実施の形態で説明した文書の表題部から得られた特徴要素に基づいてクラスタリングを行う例の他、たとえば、ＵＲＬアドレス（たとえば、http://を取り除いた部分を使用する）、更新日時（単純な時間または最近１カ月以内の更新日時）、ファイルサイズ（web ページ本文のバイトサイズなど）を用いてクラスタリングすることもできる。また、これらは、単独で用いてクラスタリングするようにしてもよく、幾つかを組み合わせてもよい。これらのどれを用いるかは、最初にメニューなどで選択項目を選ぶことで可能となる。また、選んだ項目が無い場合には、他の項目を代用する。たとえば、タイトルを選んだ場合、web ページにタイトルが無い場合には、ＵＲＬアドレスを代用する。

そして、いずれかの方法によってクラスタリングされたのち、そのクラスタリング結果に対し、前述の実施の形態で説明したような処理、すなわち、それぞれのクラスタに含まれる文書の共通性を判断してそれぞれのクラスタ同志を統合するか否かを決めるという処理を施すことによってもクラスタマージを行うことができる。

たとえば、ＵＲＬによってクラスタリングする場合について説明すれば、あるＵＲＬ（これをＵＲＬ１とする）のクラスタと、あるＵＲＬ（これをＵＲＬ２と

する）のクラスタに分類されたとし、ＵＲＬ１のクラスタには文書Ｄ１，Ｄ２，Ｄ３，Ｄ４が存在し、ＵＲＬ２のクラスタには文書Ｄ２，Ｄ３，Ｄ４，Ｄ５が存在したとする。この場合、これら２つのクラスタには、共通する文書として文書Ｄ２，Ｄ３，Ｄ４が含まれることになり、この共通する文書数と合計の文書数との関係から、ＵＲＬ１のクラスタとＵＲＬ２のクラスタを統合するか否かを決める。

また、クラスタマージするか否かの判断は、前述の実施の形態では、対象となるクラスタに含まれる合計の文書数で共通の文書数を割って得られる割合（％）で表し、その値が予め設定されたしきい値（％）と比較することによって行ったが、これに限られるものではなく、たとえば、共通する文書の個数を数え、その個数とそれぞれのクラスタに含まれる文書数との関係からマージするかしないかを決めるようにすることも可能である。

また、前述の実施の形態では、文書Ｄ１，Ｄ２，・・・，Ｄ７は、それぞれが独立した文書であって、それぞれ独立した文書を分類する場合について説明したが、ある１つの文書を幾つかのコンテンツに分けて、それぞれのコンテンツ（ここでいうコンテンツとは文書の中の意味的なまとまりを指す）を分類する場合にも適用できる。ここで抽出されるコンテンツは、各表題部ごとに切り分けられて得られる文書の中の意味的なまとまりであるとする。

たとえば、図２で示した文書Ｄ１，Ｄ２，・・・，Ｄ７が集まって１つの文書が構成されていると仮定すれば、文書Ｄ１，Ｄ２，・・・，Ｄ７をそれぞれコンテンツとみなすことができる。これらをコンテンツとすれば、それぞれのコンテンツは、表題部Ｔ１，Ｔ２，・・・，Ｔ７と本文Ａ１，Ａ２，・・・，Ａ７から構成されたものとなる。

このように、１つの文書を複数のコンテンツに分けて考えた場合、それぞれのコンテンツをクラスタリングし、そのクラスタリング結果をクラスタマージする場合にも同様に適応できる。

さらに、本実施形態で説明したクラスタリング対象文書は、たとえば、汎用の検索サービスで検索された複数の文書をクラスタリング対象文書として考えることもできる。この場合、検索された多数の文書に対してクラスタリング処理を行

い、そのクラスタリングされた結果についてクラスタマージ処理を行う。

また、以上説明した本実施形態の文書分類処理を行う処理プログラムは、フロッピィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

（第２の実施形態）

クラスタマージ後のクラスタをユーザに提示する際、単に、クラスタマージ処理結果が提示されたとすると、ユーザ側からみたとき、どのようなクラスタマージ処理がなされて統合されたのかといったクラスタマージ処理内容、すなわち、そのクラスタマージによって得られた新たなクラスタは、もともとどのクラスタとどのクラスタがどの程度の関連性があるから統合されたのかといった内容がわかりにくくになることがある。

そこで、本実施形態においては、内容に関連性のある複数のクラスタを統合するクラスタマージ処理がなされたあと、そのクラスタマージ処理されて得られたら新たなクラスタを表示する際、その新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性があるから統合されたのかといったクラスタマージ処理内容がわかるように表示している。

以下、本発明の第２の実施形態について詳細に説明する。

また、第２の実施形態では、文書分類の手法として、前述したように、それぞれの文書の表題部（タイトル）を検出して、その表題部から特徴要素を抽出して、抽出された特徴要素に基づいて文書を分類する手法を用いるものとする。

図９は、第２の実施形態を示すもので、大きく分けると、それぞれの文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部９１と、このクラスタリング部９１によって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ部９２と、このクラスタマージ部２によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわか

るように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ処理内容生成部９３と、その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力部９４とを有した構成となっている。

クラスタリング部９１は、文書記憶部９１１、文解析部９１２、特徴要素抽出部９１３、特徴テーブル作成部９１４、文書分類部９１５、分類結果記憶部９１６を有している。

文書記憶部９１１はこの場合、多数の文書データをデータベースとして持つものである。ここでは、たとえば、図１０に示すような文書群を分類する場合を説明する。図１０に示される文書群は、それぞれが独立した文書Ｄ１，Ｄ２，・・・，Ｄ７を有し、これらの文書Ｄ１，Ｄ２，・・・，Ｄ７は表題部Ｔ１，Ｔ２，・・・，Ｔ７と、それに対する本文Ａ１，Ａ２，・・・，Ａ７を持っているものとする。

文解析部９１２は文書記憶部９１１に記憶されている文書を文解析し、それぞれの文書の表題部を検出する。この文解析部９１２が行う表題部の検出は、具体的には次のようにして行う。

まず、第１の方法として、文書構造様式によって表題と規定される部分があればその部分を表題部とする。また、第２の方法として、文書構造様式によって、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とする。また、第３の方法として、定められた数の文または単語を文書先頭より抽出し、その抽出した部分を表題部とする。さらには、これら第１、第２、第３の方法を順次行い、第１の方法を行ったとき、表題と規定されている部分があればその部分を表題部とし、表題と規定される部分が存在しなければ、第２の方法を行い、標準より大きな文字で表示する指定がなされている部分があれば、その部分を表題部とし、標準より大きな文字で表示する指定がなされていなければ、第３の方法を行って表題部を検出する。

特徴要素抽出部９１３は、文解析部９２で検出されたそれぞれの文書の表題部の中から特徴要素を抽出する。

特徴テーブル作成手段９１４は、前記表題部から抽出された特徴要素とそれぞれの文書との関係を示す特徴テーブルを作成する。なお、この特徴テーブルの具

体的な内容については後述する。

文書分類部９１５は、前述の特徴テーブルの内容を参照し、文書Ｄ１，Ｄ２，・・・，Ｄ７を意味的に共通性のある複数のクラスタに分類する。つまり、文書Ｄ１，Ｄ２，・・・，Ｄ７の表題部に存在する特徴要素に基づいて、共通する特徴要素を持つ処理対象文書を１つのまとまりとし、そのまとまりを１つのクラスタとする。なお、この文書分類部９１５は同義特徴辞書（図示せず）を有し、共通する特徴要素を持つ処理対象文書を１つのまとまりとする処理を行う際、共通する特徴要素であるか否かの判断を、その同義語辞書を用い同義語が有るか否かにより行い、同義語が存在する場合にはそれを同じクラスタとする処理を行うことも可能である。

分類結果記憶部９１６は、文書分類部９１５によって分類された内容を記憶する。

クラスタマージ部９２は、複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合する処理を行うものであるが、その具体的な処理については後述する。

クラスタマージ処理内容生成部９３は、クラスタマージ部９２で判断されたクラスタ間の関連性の高さを示す値（後述する）を用い、その値を予め設定されたしきい値（後述する）と比較して関連性の高さを判断する関連性判断部９３１と、この関連性判断部９３１によるクラスタ間の関連性の高さに基づいて、どのようなクラスタがどのような関連性を有して統合されたかがわかるように、それぞれのクラスタ名の表示の仕方を決めるクラスタ名表示内容決定部９３２とを有し、その具体的な処理内容については後述する。

また、分類結果出力部９４は、出力制御部９４１と表示部９４２を有し、本発明による文書分類結果を出力する。

このような構成において、本発明の文書分類処理について説明する。本発明が行う概略的な文書分類処理は、図１１のフローチャートに示すように、処理対象となる多数の文書を意味的に共通性を有する複数のクラスタに分類し（ステップ１１Ｓ１）、これにより分類された複数のクラスタ間で各々のクラスタに含まれ

る文書に基づいて、それぞれのクラスタ間の関連性を判断する（ステップ１１Ｓ２）。そして、一定以上の関連性を有する少なくとも２つのクラスタを統合する（ステップ１１Ｓ３）。その後、クラスタマージされて得られた新たなクラスタは、どのようなクラスタがどのような関連性を有して統合されたかがわかるようなクラスタマージ内容を生成する。具体的には、クラスタマージされたクラスタ間の関連性の高さを判定し（ステップ１１Ｓ４）、その関連性の高さに基づいて、統合される前の個々のクラスタに関する情報がわかるような表示内容、すなわち、クラスタマージによって得られた新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかがわかるような表示内容を生成する（ステップ１１Ｓ５）。以下、具体例を参照して詳細に説明する。

ここでは、図１０で示した文書Ｄ１，Ｄ２，・・・，Ｄ７を分類する例について説明する。この実施の形態では、それぞれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理された結果についてクラスタマージ処理を行う。まず始めに、表題部から特徴要素を抽出し、その抽出された特徴要素に基づいて行われるクラスタリング処理（クラスタリング部１が行う処理）について説明する。

これらの文書Ｄ１，Ｄ２，・・・，Ｄ７は、文解析部１２にて表題部が検出される。たとえば、文書Ｄ１については表題部Ｔ１が検出され、文書Ｄ２については表題部Ｔ２が検出され、文書Ｄ３については表題部Ｔ３が検出されるというように、それぞれの文書Ｄ１，Ｄ２，・・・，Ｄ７の表題部Ｔ１，Ｔ２，・・・，Ｔ７が検出される。

そして、特徴要素抽出部９１３によって、それぞれの表題部に存在する特徴要素が抽出されたのち、特徴テーブル作成部９１４により、それぞれの特徴要素とその特徴要素を表題部に含む文書との関係を示す特徴テーブルが作成される。この特徴テーブルの例を図１２に示す。なお、ここでは、文書数が３つ以上取り出される特徴要素とその特徴要素を含む文書との関係を示し、特徴テーブル内に示される数値は、その特徴要素が各文書の表題部に幾つ含まれるているかの数を示している。たとえば、「用紙」という特徴要素は、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７のそれぞれの表題部に、それぞれ１個ずつ含まれていることを示している。

図１２の特徴テーブルからもわかるように、表題部に「用紙」という特徴要素を含む文書は、文書Ｄ１，Ｄ４，Ｄ６，Ｄ７であり、また、表題部に「カセット」という特徴要素を含む文書は、文書Ｄ１，Ｄ４，Ｄ７であり、さらに、表題部に「増設」という特徴要素を含む文書は、文書Ｄ２，Ｄ３，Ｄ５，Ｄ７である。なお、図１０において、これら各特徴要素部分にはアンダーラインが施されている。

そして、文書分類部９１５はこのような特徴テーブルを参照して、それぞれの特徴要素ごとの文書クラスタ分けを行う。その分類結果を図１３に示す。なお、このようなクラスタに分類する際、前述したように、共通する特徴要素であるか否かの判断を、同義語辞書を用い同義語が有るか否かによっても行い、同義語が存在する場合にはそれを同じ文書クラスタとする処理を行うことも可能である。たとえば、「用紙」と「印刷紙」の両方が特徴要素として抽出されたとすれば、これらの特徴要素を表題部に含む文書は同じクラスタとするなどという処理を行う。

このような分類結果は分類結果記憶部９１６に格納される。図１３に示される分類結果において、たとえば、「用紙」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれる）について見れば、図１０の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ６は印刷された後の用紙の汚れについての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７はどれも用紙に関する内容であり、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

また、「カセット」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ７が含まれる）について見れば、図１０の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７にはどれも用紙をセットすることに関する内容が含まれており、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

また、「増設」で分類されたクラスタ（文書Ｄ２，Ｄ３，Ｄ５，Ｄ７が含まれる）について見れば、図１０の文書内容からもわかるように、文書Ｄ２はメモリの増設についての内容であり、文書Ｄ３はインタフェースカードの増設についての内容であり、文書Ｄ５はハードディスクの増設についての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ２，Ｄ３，Ｄ５，Ｄ７はどれも何かを増設する場合についての内容であり、１つのクラスタとして分類されて何等問題のないものとなり、その分類結果は適切であるといえる。

このような適切な分類が行える理由としては、それぞれの文書の表題部から特徴要素を抽出し、その特徴要素に基づいて文書を分類しているからである。つまり、文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることが多い。したがって、文書の表題部に含まれる特徴要素を用いて分類を行うことにより、分類結果が散漫になることが少なく、また、ノイズクラスタが生成される率も少なくすることができる。また、各文書の表題部は、その文書の作成者がその文書の主旨を表す内容を表現していることから、文書の制作者側の視点による分類が得られる。

そして、分類が行われた後、ユーザによって、たとえば、「用紙」についてのクラスタの選択指示が出されたとすると、そのクラスタに属する文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が文書記憶部１１から読み出されて表示部３２に表示される。なお、このときの表示内容としては、前述したように、文書番号や文書名のみでもよく、さらには、その文書内容を表示させるようにしてもよい。

ところで、本発明実施形態においては以上のようにクラスタリング処理した結果について、さらに、クラスタマージ部２によってクラスタマージ処理を行う。

すなわち、図１３に示す分類結果において、特徴要素である「用紙」と「カセット」について見ると、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれ、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７に存在することがわかる。

このように、「用紙」のクラスタと「カセット」のクラスタには、共に文書Ｄ１，Ｄ４，Ｄ７が共通して存在している。これは、「用紙」という特徴要素と「カ

セット」という特徴要素は相互に関連した状態で用いられることが多いことを意味している。たとえば、文書Ｄ１，Ｄ４，Ｄ７の表題部または本文のなかに「用紙カセット」という用語が用いられている。つまり、これらの文書Ｄ１，Ｄ４，Ｄ７は共通性の高い文書であり、これら文書Ｄ１，Ｄ４，Ｄ７は同じクラスタに分類した方がより好ましいと考えられる。

これを実現するために、特徴要素に基づいてクラスタリングしたあと、そのクラスタリング結果に対しクラスタマージ処理を施す。

このクラスタマージ処理について以下に説明する。まず始めに、図１３の分類結果とは関係なく一般的な例について図１４を参照しながら説明する。

今、２つのクラスタＣ１，Ｃ２があるとする。クラスタＣ１として５個の文書Ｄ１，Ｄ２，Ｄ３，Ｄ４，Ｄ８が抽出され、クラスタＣ２には６個の文書Ｄ３，Ｄ４，Ｄ５，Ｄ６，Ｄ７，Ｄ８が抽出されたとする。

ここで、２つのクラスタＣ１，Ｃ２に共通している文書は、文書Ｄ３，Ｄ４，Ｄ８である。この実施の形態では、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基に、それぞれのクラスタ間の関連性を判断してクラスタマージ処理を行う。

具体的には、複数のクラスタとして、ある２つのクラスタに共通している文書数が２つのクラスタに存在する合計の文書数に対しどのくらいの割合かを計算し、その計算結果が予め定めたしきい値以上かどうかによってマージするか否かを決める。

たとえば、この場合、２つのクラスタＣ１，Ｃ２に存在する文書数の合計は１１個であり、両者に共通する文書数は３個である。これらから合計の文書数に占める共通する文書数の割合（％）を計算し、その結果からマージするか否かを決定する。この割合（％）を求める際、合計の文書数で共通する文書数を単純に割り算してそれに１００を掛けて求めてもよいが、共通する文書数に任意に設定される係数を掛け算したものを合計の文書数で割り算してそれに１００を掛けて求めるようにしてもよい。

一例として、クラスタＣ１に存在する文書数をα１、クラスタＣ２に存在する文書数をα２とし、両者に共通する文書数をβとした場合、たとえばβに係数と

してたとえば２を掛けて、２β／（α１＋α２）×１００を計算し、その値（％）が予め設定されたしきい値ＴＨ（％）と比較して、上式による計算結果がしきい値ＴＨ以上であればマージするというようなことを行う。図１４で示した例について考えれば、２βは２×３＝６個、α１＋α２は５＋６＝１１個であるので、この場合、約５５％と求められる。ここで、しきい値ＴＨが仮に７０％と設定されているとすれば、計算結果（５５％）はしきい値ＴＨ（７０％）より小さいので、クラスタＣ１とクラスタＣ２はマージしないとする。なお、係数は任意に設定されるもので、計算結果で得られる数値（％）がしきい値と比較し易いような値となるように適当に設定されるものであり、この場合は係数を２としたが、係数を１としても特に問題はない。

ここで、図１３で示した分類結果を例にして説明すれば、図１３の場合、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７の３つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ１，Ｄ４，Ｄ７の３つの文書であり、これを合計の文書数に対する割合（％）で考える。

これを前述した計算式によって計算する。図１３の分類結果の場合、合計の文書数（α１＋α２）は、４＋３＝７となり、共通の文書数は３で２βは６となる。したがって、この場合、約８６％という高い値が得られる。これは、設定されたしきい値（ここでは７０％としている）よりも高いので、この「用紙」のクラスタと「カセット」のクラスタはマージして１つのクラスタとするということになる。

同様に考えて、図１３の「用紙」のクラスタと「増設」のクラスタとをマージするか否か、「カセット」のクラスタと「増設」のクラスタとをマージするか否かについて判断する。

まず、「用紙」のクラスタと「増設」のクラスタについては、「用紙」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを上式を用いて計算すると、この場合、２５％という結果が得られ、これは、しきい値（７０％）よりも低い

値であるので、この場合は、両者はマージしないとする。

また、「カセット」のクラスタと「増設」のクラスタについては、「カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ７の３つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを上式を用いて計算すると、この場合、約２８％という結果が得られ、これは、しきい値（７０％）よりも低い値であるので、この場合は、両者はマージしないとする。

このようにして、それぞれのクラスタに対し２つのクラスタごとにそれぞれマージするか否かを判断する。この図１３の分類結果についてマージするか否かの処理を行ったあとの分類結果（マージ処理後の分類結果という）が図１５である。図１５によれば、「用紙」と「カセット」が「用紙＋カセット」という１つのクラスタに分類され、そのクラスタに属する文書は文書Ｄ１，Ｄ４，Ｄ６，Ｄ７ということになる。また、「増設」についてはそのまま単独のクラスタを構成する。

図１５に示されるクラスタマージ処理後の分類結果において、たとえば、「用紙＋カセット」で分類されたクラスタ（文書Ｄ１，Ｄ４，Ｄ６，Ｄ７が含まれる）について見れば、図１０の文書内容からもわかるように、文書Ｄ１は用紙カセットについての内容であり、文書Ｄ４は用紙設定についての内容であり、文書Ｄ６は印刷された後の用紙の汚れた場合にはどのようにするかについての内容であり、文書Ｄ７は用紙カセットの増設についての内容である。

このように、これらの文書Ｄ１，Ｄ４，Ｄ６，Ｄ７はどれも用紙やカセットに関する内容であり、１つのクラスタとして分類されて何等問題のないものとなり、むしろ、「用紙＋カセット」を１つのクラスタとした方がよい分類結果であるといえる。

このように、始めにそれぞれの文書の表題部から特徴要素を抽出し、その抽出された特徴要素に基づいてクラスタリング処理を行い、かつ、そのクラスタリング処理されて得られたそれぞれのクラスタに対し、２つづつのクラスタの組み合わせについてクラスタマージ処理を行うことによって、より適切なクラスタリングが行える。

また、以上のようにして２つのクラスタごとに１回目のクラスタマージ処理が

終了し、図１５のようなクラスタマージ処理後の分類結果が得られると、今度は、そのクラスタマージ処理後の分類結果について、２回目のクラスタマージ処理を行う。つまり、図１５の１回目のクラスタマージ処理後の結果で考えた場合、「用紙＋カセット」のクラスタと「増設」のクラスタについてクラスタマージ処理を行う。この場合、「用紙＋カセット」のクラスタと「増設」のクラスタについては、「用紙＋カセット」のクラスタには文書Ｄ１，Ｄ４，Ｄ６，Ｄ７の４つの文書が存在し、「増設」のクラスタには文書Ｄ２，Ｄ３，Ｄ５，Ｄ７の４つの文書が存在する。そして、２つのクラスタに共通する文書は文書Ｄ７のみであり、これを合計の文書数に対する割合（％）で考えると、共通する文書数１に定数２を掛けたものを合計の文書数８で割り算し、それに１００を掛けると、２５％という結果が得られ、これは、しきい値（７０％）よりも低い値であるので、この場合は、両者はマージしないとする。

このようにして、２つのクラスタ間で１回目のクラスタマージ処理が終了した後、その１回目のクラスタマージ処理に新たな２つのクラスタ間で２回目のクラスタマージ処理を行い、その２回目のクラスタマージ処理が終了した後、その２回目のクラスタマージ処理後に新たな２つのクラスタ間で３回目のクラスタマージ処理を行うというクラスタマージ処理を順次行い、新たなクラスタが生成されなくなるまで（クラスタマージが起こらなるまで）その処理を繰り返す。

また、これまでの説明では、２つのクラスタ間でクラスタマージ処理を行う例について説明したが、クラスタマージ処理は３つ以上のクラスタの組み合わせについても可能である。この場合、１回のクラスタマージ処理によって３つ以上のクラスタ間でクラスタマージ処理を行い、さらに、これによって幾つかのクラスタに分類された結果についてクラスタマージが起こらなくなるまで、順次、クラスタマージ処理を行うことも可能である。なお、３つ以上のクラスタについてクラスタマージするか否かを判断する場合、前述したように、それぞれのクラスタに存在する合計の文書数に対する共通の文書数の割合（％）で考えることができる。

以上のようにして、図９に示したクラスタマージ部９２によるクラスタマージ処理が終了すると、次に、クラスタマージ処理内容生成部９３がそのクラスタマ

ージ結果に対し、クラスタマージされたクラスタ間の関連性の高さを判定し、その関連性の高さに基づいて、統合される前の個々のクラスタに関する情報がわかるような表示内容、すなわち、クラスタマージによって得られた新たなクラスタは、どのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかがわかるような表示内容を生成する。以下、このクラスタマージ処理内容生成部９３が行う処理について説明する。

この実施の形態では、クラスタマージ部９２によって得られた関連性の高さとしての関連性の度合い（％）の値が、前述したしきい値ＴＨよりずっと大きい値であるか、しきい値ＴＨに近い値であるかによって、そのクラスタマージされたクラスタ間の関連性の高さを関連性判断部９３１によって判断する。具体的には、前述のしきい値ＴＨに対し、それよりも高い値（％）のしきい値ＴＨ１を設定し、クラスタマージ部９２によって得られた関連性の度合い（Ｋで表す）が、Ｋ≧ＴＨ１であれば、クラスタ同志の関連性はきわめて大きく殆ど同じ内容であると判断する。一方、クラスタマージ部９２によって得られた関連性の度合いＫが、ＴＨ１＞Ｋ≧ＴＨであれば、少し似ている程度と判断する。

今、Ｋ≧ＴＨ１である場合、すなわち、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタ同志の関連性がきわめて高い場合は、次のような処理を行う。

これを図１５の例で説明すれば、クラスタマージされた新たなクラスタの特徴要素は、「用紙＋カセット」である。この「用紙＋カセット」のクラスタは、図１３に示す用紙のクラスタとカセットのクラスタをクラスタマージした結果である。

このそれぞれのクラスタにクラスタ名を付けるとすれば、特徴要素が「用紙」であるクラスタを「用紙クラスタ」、特徴要素が「カセット」であるクラスタを「カセットクラスタ」といように表すことができ、それぞれのクラスタ名を以下では、単に、「用紙」、「カセット」と表記する。

ここで、クラスタマージ結果である「用紙＋カセット」のクラスタは、クラスタマージ部９２による前述の計算によって、８６％という値が得られている。ここで、関連性判断部９３１において、関連性を判断する際に設定されたしきい値

ＴＨ１が８０％と設定されているとすれば、この場合、クラスタマージ部９２によって得られた関連性の度合いＫは、Ｋ≧ＴＨ１であるので、用紙クラスタとカセットクラスタの関連性はきわめて大きく殆ど同じ内容であると判断できる。

このように、クラスタマージ部９２によって得られた関連性を示す値Ｋが、Ｋ≧ＴＨ１である場合には、クラスタマージされたそれぞれのクラスタ同志の関連性はきわめて大きく、殆ど同じ内容のクラスタであると判断でき、それぞれのクラスタの名称を、連続的に表示する。たとえば、上述の「用紙クラスタ」と「カセットクラスタ」の例では、それらのクラスタ名である「用紙」と「カセット」をくっつけて「用紙カセット」などと表記してそれを表示する。

これは、いわゆるＡＮＤ形式の表記の仕方であり、クラスタ名をくっつけて表記しても差し支えないような場合である。この例では、クラスタマージされて得られた新たなクラスタのクラスタ名を「用紙カセット」とすることになるが、この場合は、クラスタマージされて得られた新たなクラスタは、その新たなクラスタを構成する用紙クラスタとカセットクラスタに含まれるそれぞれの文書内容（図１０参照）から見て、新たなクラスタ名を「用紙カセット」として何等差し支えないものである。

図１６はこのような処理を行ったあとの表示例を示すもので、この図１６では、クラスタマージされた新たなクラスタのクラスタ名としての「用紙カセット」と、その新たなクラスタに含まれる文書として、ここでは、図１０で示されたそれぞれの文書（文書Ｄ１，Ｄ４，Ｄ６．Ｄ７）のそれぞれの表題（タイトル）名が表示されている。

また、このように、それぞれのクラスタ名を、連続的に表示する方法の他に、図１７に示すように、それぞれのクラスタ対応のクラスタ名である「用紙」と「カセット」を、それぞれのクラスタ名ごとに改行して縦に並べて表記するようにしてもよい。

このように、それぞれのクラスタの名称を縦に並べると、言語的なつながりが気にならなくなり、違和感を与えない効果がある。この実施の形態で用いている「用紙」と「カセット」は、連続して「用紙カセット」としても何等問題ないが、場合によっては、違和感を持つ場合もある。たとえば、これまでの説明とは全く

関係のない例として、クラスタマージされた得られた新たなクラスタに含まれるそれぞれのクラスタ名が、仮に、「製品」、「使用」、「概要」であったとする。このようなクラスタ名を上述のように、連続して横に一列に並べると「製品仕様概要」となる。これでも意味が全く不明というものではないが、言語的に少し違和感が生じる。このような場合、本来は、言語処理を行って、「製品仕様の概要」というようにすればよいが、そのような言語処理は複雑で時間を要する。
したがって、このような場合、図１７と同様に、「製品」、「使用」、「概要」を１つづつ縦に並べると違和感を与えることがなくなる。また、縦に並べることで、実際に表示したときに、横並び一列での表示に比べ、クラスタマージされたクラスタ名の数が多くても、横方向にむやみに伸びることがないので見易くなるという効果もある。

このように、クラスタマージ部９２によって得られた関連性を示す値Ｋが、Ｋ≧ＴＨ１であって、クラスタマージされて得られた新たにクラスタに含まれるクラスタのクラスタ名をＡＮＤ形式の表記とし、クラスタ名を横一列に並べた表記の仕方で表示するか、あるいは、各クラスタ対応のクラスタ名称ごとに改行して縦に並べる表記の仕方で表示する。

これによって、クラスタマージされて得られた新たなクラスタは、どのようなクラスタがどのような関連性を有して統合されたかということが、そのクラスタマージされた新たなクラスタ名を見るだけでわかる。たとえば、図１６や図１７の例では、元のクラスタは「用紙」というクラスタと「カセット」というクラスタが統合されてできたクラスタであり、しかも、その関連性はきわめて高く同じような内容の文書を持ったクラスタであるということがわかる。

次に、ＴＨ１＞Ｋ≧ＴＨである場合、すなわち、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタの関連性の度合いは、殆どがオーバーラップするほどでもないが同じ文書を幾つか含んでいるといった場合の処理について説明する。

このように、クラスタマージ部９２によって得られた関連性を示す値Ｋが、ＴＨ１＞Ｋ≧ＴＨである場合には、それぞれのクラスタの名称を、いわゆるＯＲ形式の表記の仕方で行う。

たとえば、前述の「製品」、「使用」、「概要」の例で説明すれば、この場合、「製品」、「使用」、「概要」を連続的な表示ではなく、たとえば、「製品・使用・概要」というように、それぞれの名称間に区切りの記号を挿入して表示する。このような区切りの記号がある場合にはＯＲ的な内容であることを予めユーザに報知しておけば、それを見たユーザはそのクラスタマージされて得られた新たなクラスタには、「製品」、「使用」、「概要」といった内容を持った文書が幾つか含まれているというように理解できる。なお、このＯＲ形式の表記の仕方を行う場合、クラスタ名の間に挿入する記号は上述したような「製品・使用・概要」の例に限られるものではなく、たとえば、クラスタ名の間に「／」を挿入して「製品／使用／概要」ようにしてもよい。

また、クラスタマージされて得られた新たなクラスタに含まれる幾つかのクラスタの関連性に、Ｋ≧ＴＨ１と、ＴＨ１＞Ｋ≧ＴＨが混在するような場合もある。このような場合には、それぞれの関連性の度合いがわかるように、ＡＮＤ形式とＯＲ形式に分けて表記する。

さらに、クラスタマージされたそれぞれのクラスタ同志が包含関係にあるような場合もある。たとえば、あるクラスタが「製品」に関するクラスタであり、あるクラスタのクラスタ名が「テレビ」、あるクラスタのクラスタ名が「ラジオ」、あるクラスタのクラスタ名が「ビデオ」であって、これらのクラスタがクラスタマージされたとする。このとき、「テレビ」のクラスタ、「ラジオ」のクラスタ、「ビデオ」のクラスタが「製品」のクラスタに包含されるものであって、しかも、それぞれのクラスタ同志の関連性の度合いがＴＨ１＞Ｋ≧ＴＨの関係であったとすれば、「製品・（テレビ・ラジオ・ビデオ）」というような表記の仕方で表示する。これは、「製品」、「テレビ」・「ラジオ」・「ビデオ」はそれぞれがＯＲ的な関係にあり、しかも、「テレビ」・「ラジオ」・「ビデオ」が括弧でくくられていることから、これら「テレビ」・「ラジオ」・「ビデオ」の各クラスタは「製品」に包含されるクラスタであることを意味している。

このように、クラスタマージ処理がなされて得られた新たなクラスタのクラスタ名を見るだけで、どのようなクラスタがどの程度の関連性を有して統合されたのかを容易に知ることができる。

なお、本実施形態は上記内容に限定されるものではなく、本実施形態の要旨を逸脱しない範囲で種々変形実施可能となるものである。たとえば、前述の実施の形態では、図１３に示すような分類結果を得るための特徴要素を各文書の表題部から得るようにして、表題部から得られた特徴要素に基づいたクラスタリングを行う例について説明したが、本発明においては、複数の文書をクラスタリングする手法は、特に限定されるものではない。

複数の文書をクラスタリングする手法としては、前述の実施の形態で説明した文書の表題部から得られた特徴要素に基づいてクラスタリングを行う例の他に、たとえば、ＵＲＬアドレス（http://を取り除いた部分）、更新日時（単純な時間または最近１カ月以内の更新日時）、ファイルサイズ（web ページ本文のバイトサイズなど）を用いてクラスタリングすることもできる。また、これらは、単独で用いてクラスタリングするようにしてもよく、幾つかを組み合わせてもよい。これらのどれを用いるかは、最初にメニューなどで選択項目を選ぶことで可能となる。また、選んだ項目が無い場合には、他の項目を代用する。たとえば、タイトルを選んだ場合、web ページにタイトルが無い場合には、ＵＲＬアドレスを代用する。

そして、いずれかの方法によってクラスタリングされたのち、そのクラスタリング結果に対し、前述の実施の形態で説明したような処理、すなわち、それぞれのクラスタに含まれる文書の共通性を判断してそれぞれのクラスタ同志を統合するか否かを決めるという処理を施すことによってもクラスタマージを行うことができる。

たとえば、ＵＲＬによってクラスタリングする場合について説明すれば、あるＵＲＬ（これをＵＲＬ１とする）のクラスタと、あるＵＲＬ（これをＵＲＬ２とする）のクラスタに分類されたとし、ＵＲＬ１のクラスタには文書Ｄ１，Ｄ２，Ｄ３，Ｄ４が存在し、ＵＲＬ２のクラスタには文書Ｄ２，Ｄ３，Ｄ４，Ｄ５が存在したとする。この場合、これら２つのクラスタには、共通する文書として文書Ｄ２，Ｄ３，Ｄ４が含まれることになり、この共通する文書数と合計の文書数との関係から、ＵＲＬ１のクラスタとＵＲＬ２のクラスタを統合するか否かを決める。

また、クラスタマージするか否かの判断は、前述の実施の形態では、対象となるクラスタに含まれる合計の文書数で共通の文書数を割って得られる割合（％）で表し、その値が予め設定されたしきい値（％）と比較することによって行ったが、これに限られるものではなく、たとえば、共通する文書の個数を数え、その個数とそれぞれのクラスタに含まれる文書数との関係からマージするかしないかを決めるようにすることも可能である。

このように、個数によってクラスタマージするか否かを判断する場合、前述したしきい値は個数を用いればよい、たとえば、合計の文書数が１０個あって、共通する文書が７個以上であるときにマージするとした場合、前述のしきい値ＴＨは、たとえば７個で、ＴＨ１をたとえば９個とし、９個以上共通した文書がある場合にはＡＮＤ形式の表記の仕方での表示を行い、７個または８個の場合はＯＲ形式の表記の仕方での表示を行うというようにもできる。なお、この数値は一例であってこれに限られるものではないことは言うまでもない。これは、前述の実施の形態のなかで説明したしきい値ＴＨやＴＨ１の値についても同様のことがいえる。

また、前述の実施の形態では、文書Ｄ１，Ｄ２，・・・，Ｄ７は、それぞれが独立した文書であって、それぞれ独立した文書を分類する場合について説明したが、ある１つの文書を幾つかのコンテンツに分けて、それぞれのコンテンツ（ここでいうコンテンツとは文書の中の意味的なまとまりを指す）を分類する場合にも適用できる。ここで抽出されるコンテンツは、各表題部ごとに切り分けられて得られる文書の中の意味的なまとまりであるとする。

たとえば、図１０で示した文書Ｄ１，Ｄ２，・・・，Ｄ７が集まって１つの文書が構成されていると仮定すれば、文書Ｄ１，Ｄ２，・・・，Ｄ７をそれぞれコンテンツとみなすことができる。これらをコンテンツとすれば、それぞれのコンテンツは、表題部Ｔ１，Ｔ２，・・・，Ｔ７と本文Ａ１，Ａ２，・・・，Ａ７から構成されたものとなる。

このように、１つの文書を複数のコンテンツに分けて考えた場合、本発明はそれぞれのコンテンツをクラスタリングし、そのクラスタリング結果をクラスタマージする場合にも同様に適応できる。

さらに、本実施形態で用いられるクラスタリング対象文書は、たとえば、汎用の検索サービスで検索された複数の文書をクラスタリング対象文書として考えることもできる。この場合、検索された多数の文書に対してクラスタリング処理を行い、そのクラスタリングされた結果についてクラスタマージ処理を行う。そして、クラスタマージされて得られた新たなクラスタに含まれるそれぞれのクラスタについて前述の実施の形態で説明したように処理を行うことで、そのクラスタマージによって得られた新たなクラスタは、もともとどのクラスタとどのクラスタがどの程度の関連性を有して統合されたのかといった内容を容易に知ることができる。

また、以上説明した文書分類処理を行う処理プログラムは、フロッピィディスク、光ディスク、ハードディスクなどの記録媒体に記録させておくことができ、本発明はその記録媒体をも含むものである。また、ネットワークから処理プログラムを得るようにしてもよい。

請求の範囲

1．　複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行うことを特徴とする文書分類方法。

2．　前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタに含まれる複数の文書のうち、それぞれのクラスタに共通して含まれる文書数を基にクラスタ間の関連性を判断してクラスタマージすることを特徴とする請求項１記載の文書分類方法。

3．　前記クラスタマージ処理は、クラスタマージ処理対象となる複数のクラスタそれぞれを特徴づける特徴要素が、そのクラスタマージ処理対象となるそれぞれのクラスタに含まれる元の文書内容にどのような状態で出現するかを調べ、その出現状態に基づいてクラスタマージすることを特徴とする請求項１記載の文書分類方法。

4．　前記クラスタマージ処理は、少なくとも２つのクラスタ間で行い、一回目のクラスタマージ処理が終了すると、そのクラスタマージ処理されたクラスタ群に対し、再度のクラスタマージ処理を行い、クラスタマージが起こらなくなるまでそれを繰り返すことを特徴とする請求項１から３のいずれか１項に記載の文書分類方法。

5．　前記クラスタマージ処理を行った後は、クラスタマージを実行したことおよびクラスタマージを行った根拠を付加情報として出力することを特徴とする請求項１から４のいずれか１項に記載の文書分類方法。

6．　複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類方法において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラ

スタを統合するクラスタマージ処理を行い、

このクラスタマージ処理によって得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成し、その表示内容をユーザに提示すべき分類結果に含めて出力することを特徴とする文書分類方法。

７．　前記クラスタマージ処理内容がわかるような表示内容とは、前記統合されたそれぞれのクラスタ間の関連性の高さに基づき、当該それぞれのクラスタのクラスタ名の表示の仕方を変えた表示内容であって、それぞれのクラスタ名の表示の仕方は、前記クラスタ間の関連性の高さが予め設定された値より大きい場合には、それぞれのクラスタ名をＡＮＤ形式の表記の仕方で表示させ、前記クラスタ間の関連性の高さが予め設定された値未満である場合には、それぞれのクラスタ名をＯＲ形式の表記の仕方で表示させることを特徴とする請求項６に記載の文書分類方法。

８．　前記ＡＮＤ形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名を横方向に並べて連続的に表記するか、それぞれのクラスタ対応のクラスタ名ごとに改行して縦に並べて表記するかのいずれかで行い、前記ＯＲ形式の表記の仕方は、それぞれのクラスタ対応のクラスタ名の間に区切り記号を挿入して表記することを特徴とする請求項７に記載の文書分類方法。

９．　あるクラスタの中に包含されるようなクラスタが存在する場合には、包含されるクラスタ名を、包含するクラスタのクラスタ名に対し括弧書きの表記の仕方で表示することを特徴とする請求項７または８に記載の文書分類方法。

１０．　複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、

このクラスタリング部により得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ部と、

を有することを特徴とする文書分類装置。

１１．　複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類装置において、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング部と、

このクラスタリング部によって得られた複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ部と、

このクラスタマージ部によってクラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成するクラスタマージ内容生成部と、

その表示内容をユーザに提示すべき分類結果に含めて出力する分類結果出力手段と、

を有したことを特徴とする文書分類装置。

１２．　複数の文書を意味的に共通性を有する複数のクラスタに分類する文書分類処理プログラムを記録した記録媒体であって、その文書分類処理プログラムは、

前記複数の文書を意味的に共通性を有する複数のクラスタに分類するクラスタリング処理手順と、

これにより分類された複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞれのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理手順と、

を含むことを特徴とする文書分類処理プログラムを記録した記録媒体。

１３．　複数の文書を意味的に共通性を有する複数のクラスタに分類して出力する文書分類処理プログラムを記録した記録媒体であって、その処理プログラムは、

複数の文書を意味的に共通性を有する複数のクラスタに分類する手順と、

その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれぞ

れのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行う手順と、

クラスタマージ処理されて得られた新たなクラスタの表示を行う際、その新たなクラスタに対し、クラスタマージ処理内容がわかるように、どのようなクラスタがどのような関連性を有して統合されたかを示す表示内容を生成する手順と、

その表示内容をユーザに提示すべき分類結果に含めて出力する手順と、

を含むことを特徴とする文書分類処理プログラムを記録した記録媒体。

要約書

複数の文書をそれぞれ解析して表題部を検出する文解析部１２と、文解析部１２で検出されたそれぞれの処理対象文書の表題部から特徴要素を抽出する特徴要素抽出部１３と、表題部から抽出された特徴要素とその特徴要素を含む処理対象文書との関係を示す特徴テーブルを作成する特徴テーブル作成手段１４と、作成された特徴テーブルの内容を参照して前記処理対象文書を意味的に共通性を有する複数のクラスタに分類する文書分類部１５と、文書分類部１５により分類されたクラスタを記憶する分類結果記憶部１６と、分類結果記憶部６に記憶されたクラスタをクラスタマージ処理するクラスタマージ部２と、そのクラスタマージ処理結果を表示部３２に出力する出力制御部３１とを有した構成とする。

図 1

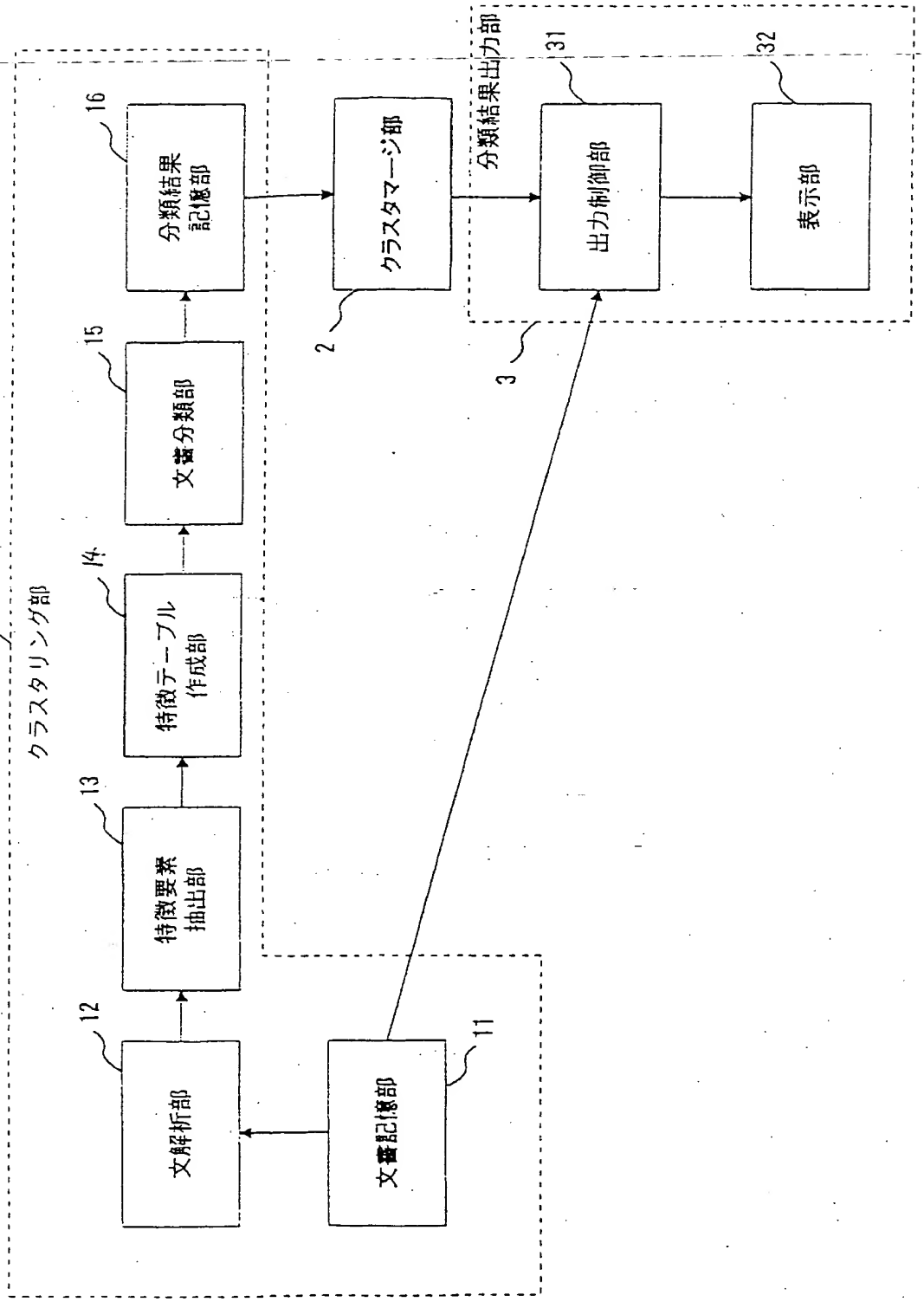
1
クラスタリング部
分類結果出力部
3
2
11
12
13
14
15
16
31
32
文書記憶部
文解析部
特徴要素抽出部
特徴テーブル作成部
文書分類部
分類結果記憶部
クラスタマージ部
出力制御部
表示部

図 2

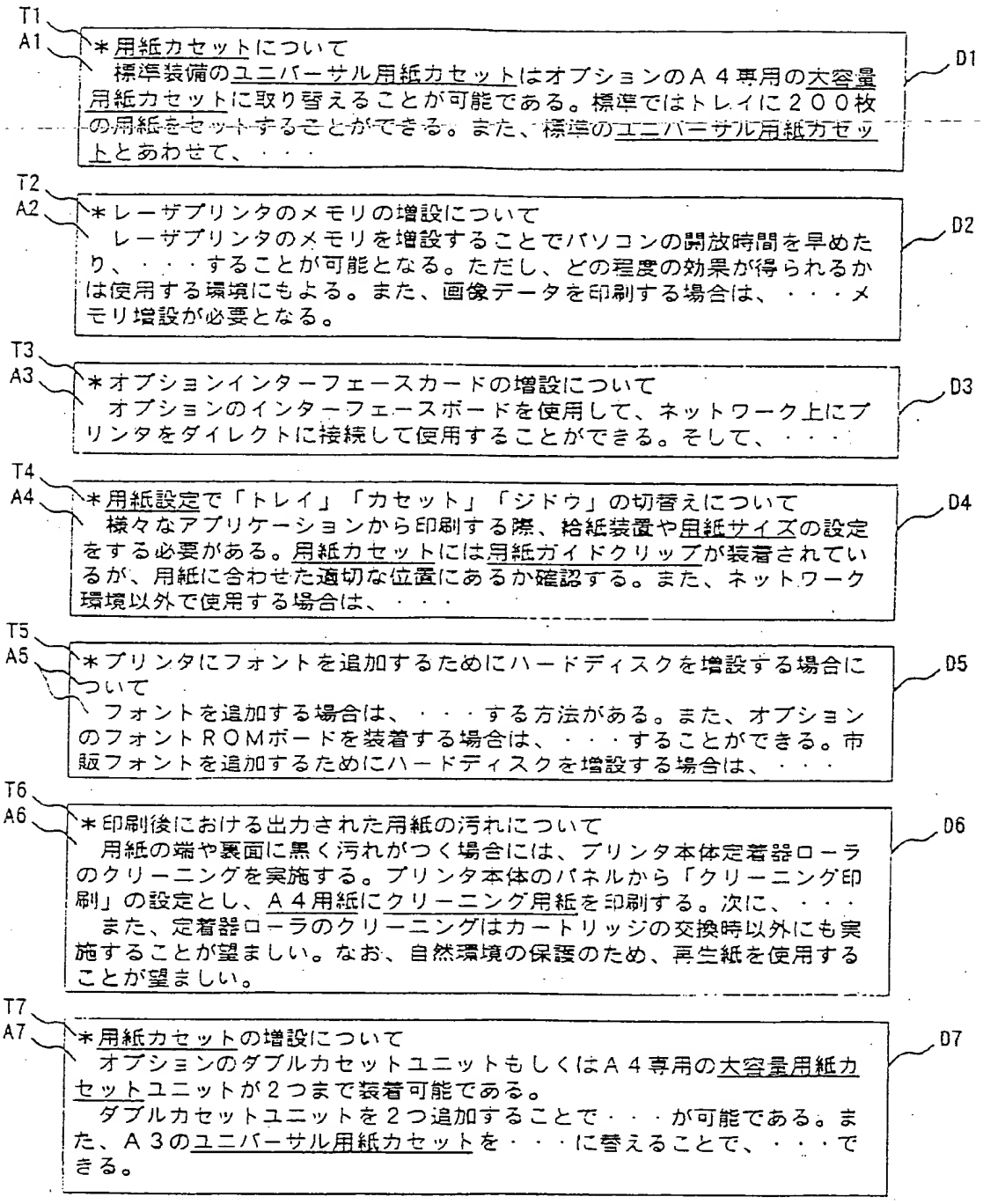
T1
A1
＊用紙カセットについて
標準装備のユニバーサル用紙カセットはオプションのＡ４専用の大容量用紙カセットに取り替えることが可能である。標準ではトレイに２００枚の用紙をセットすることができる。また、標準のユニバーサル用紙カセットとあわせて、・・・
D1
T2
A2
＊レーザプリンタのメモリの増設について
レーザプリンタのメモリを増設することでパソコンの開放時間を早めたり、・・・することが可能となる。ただし、どの程度の効果が得られるかは使用する環境にもよる。また、画像データを印刷する場合は、・・・メモリ増設が必要となる。
D2
T3
A3
＊オプションインターフェースカードの増設について
オプションのインターフェースボードを使用して、ネットワーク上にプリンタをダイレクトに接続して使用することができる。そして、・・・
D3
T4
A4
＊用紙設定で「トレイ」「カセット」「ジドウ」の切替えについて
様々なアプリケーションから印刷する際、給紙装置や用紙サイズの設定をする必要がある。用紙カセットには用紙ガイドクリップが装着されているが、用紙に合わせた適切な位置にあるか確認する。また、ネットワーク環境以外で使用する場合は、・・・
D4
T5
A5
＊プリンタにフォントを追加するためにハードディスクを増設する場合について
フォントを追加する場合は、・・・する方法がある。また、オプションのフォントＲＯＭボードを装着する場合は、・・・することができる。市販フォントを追加するためにハードディスクを増設する場合は、・・・
D5
T6
A6
＊印刷後における出力された用紙の汚れについて
用紙の端や裏面に黒く汚れがつく場合には、プリンタ本体定着器ローラのクリーニングを実施する。プリンタ本体のパネルから「クリーニング印刷」の設定とし、Ａ４用紙にクリーニング用紙を印刷する。次に、・・・
また、定着器ローラのクリーニングはカートリッジの交換時以外にも実施することが望ましい。なお、自然環境の保護のため、再生紙を使用することが望ましい。
D6
T7
A7
＊用紙カセットの増設について
オプションのダブルカセットユニットもしくはＡ４専用の大容量用紙カセットユニットが２つまで装着可能である。
ダブルカセットユニットを２つ追加することで・・・が可能である。また、Ａ３のユニバーサル用紙カセットを・・・に替えることで、・・・できる。
D7

図 3

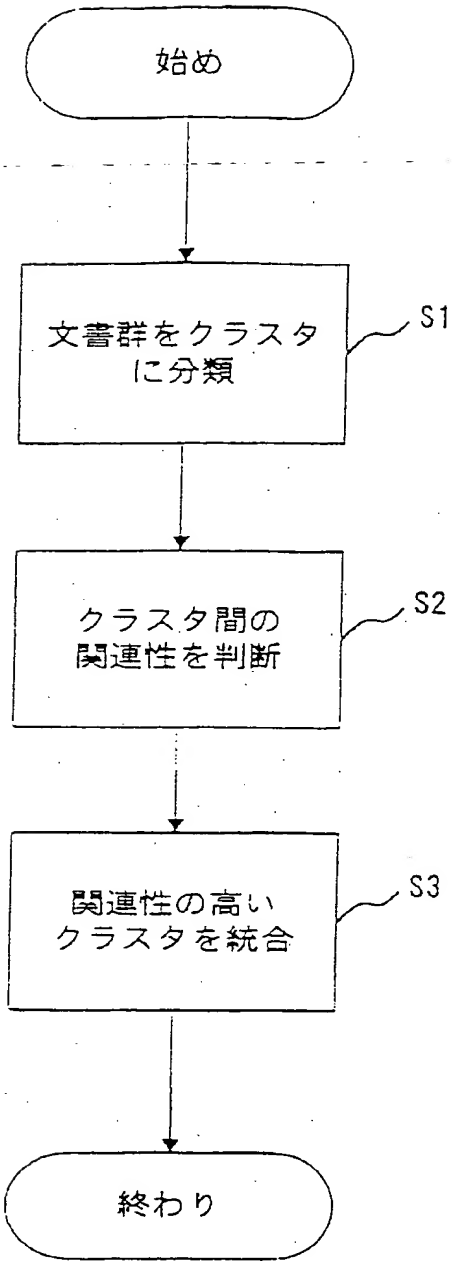

図 4

特徴要素	文書D1	文書D2	文書D3	文書D4	文書D5	文書D6	文書D7
用紙	1			1		1	1
カセット	1			1			1
増設		1	1		1		1

図 5

特徴要素	クラスタ
用紙	D1, D4, D6, D7
カセット	D1, D4, D7
増設	D2, D3, D5, D7

図 6

クラスタC1	D1, D2, D3, D4, D8
クラスタC2	D3, D4, D5, D6, D7, D8

図 7

特徴要素	クラスタ
用紙+カセット	D1, D4, D6, D7
増設	D2, D3, D5, D7

図8

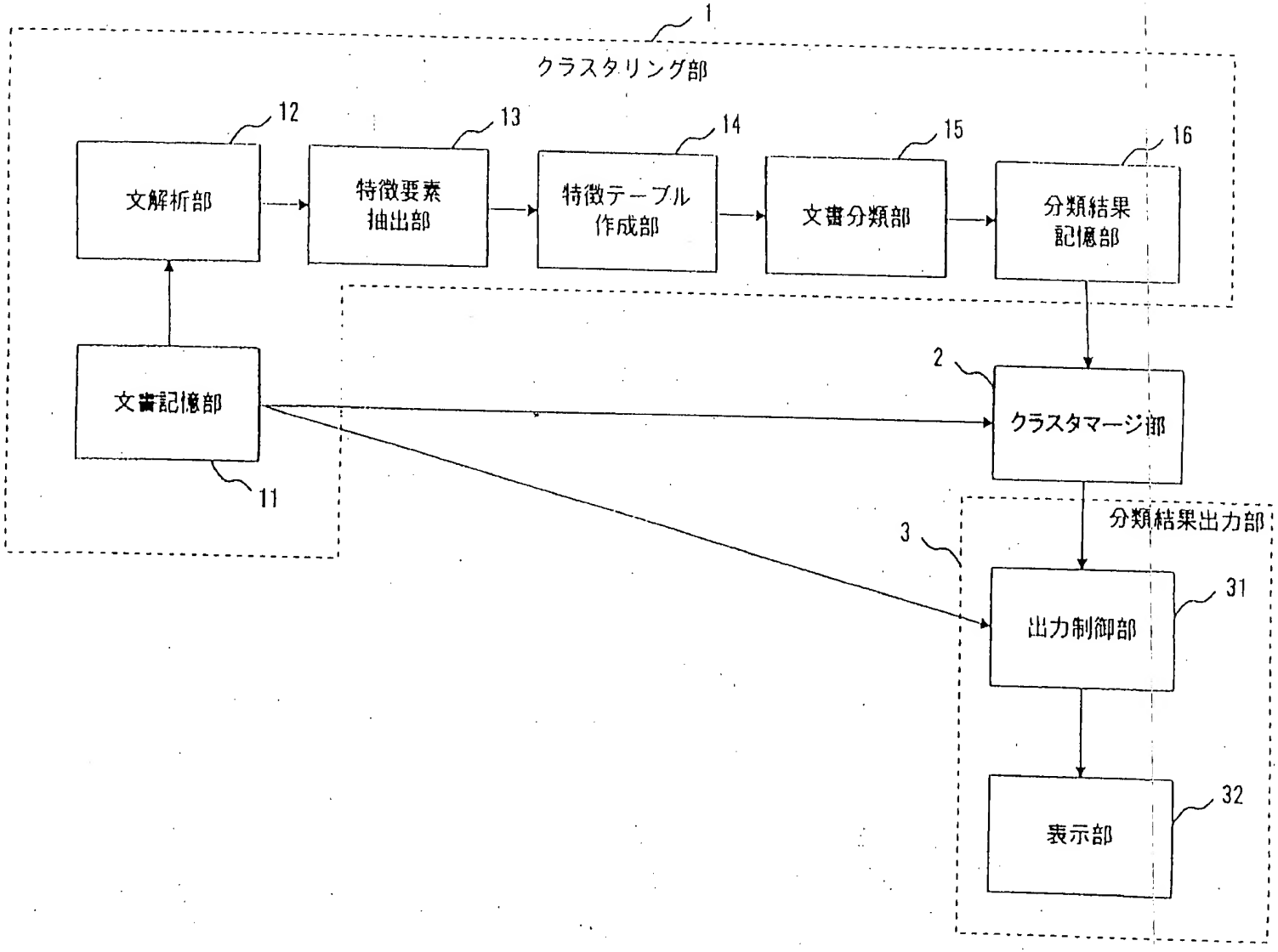
1
クラスタリング部
12
文解析部
13
特徴要素
抽出部
14
特徴テーブル
作成部
15
文書分類部
16
分類結果
記憶部
文書記憶部
11
2
クラスタマージ部
分類結果出力部
3
出力制御部
31
表示部
32

図 9

91
クラスタリング部
911
文書記憶部
912
文解析部
913
特徴要素抽出部
914
特徴テーブル作成部
915
文書分類部
916
分類結果記憶部
92
クラスタマージ部
93
クラスタマージ処理内容生成部
931
関連性判断部
932
クラスタ名表示内容決定部
94
分類結果出力部
941
出力制御部
942
表示部

図１０

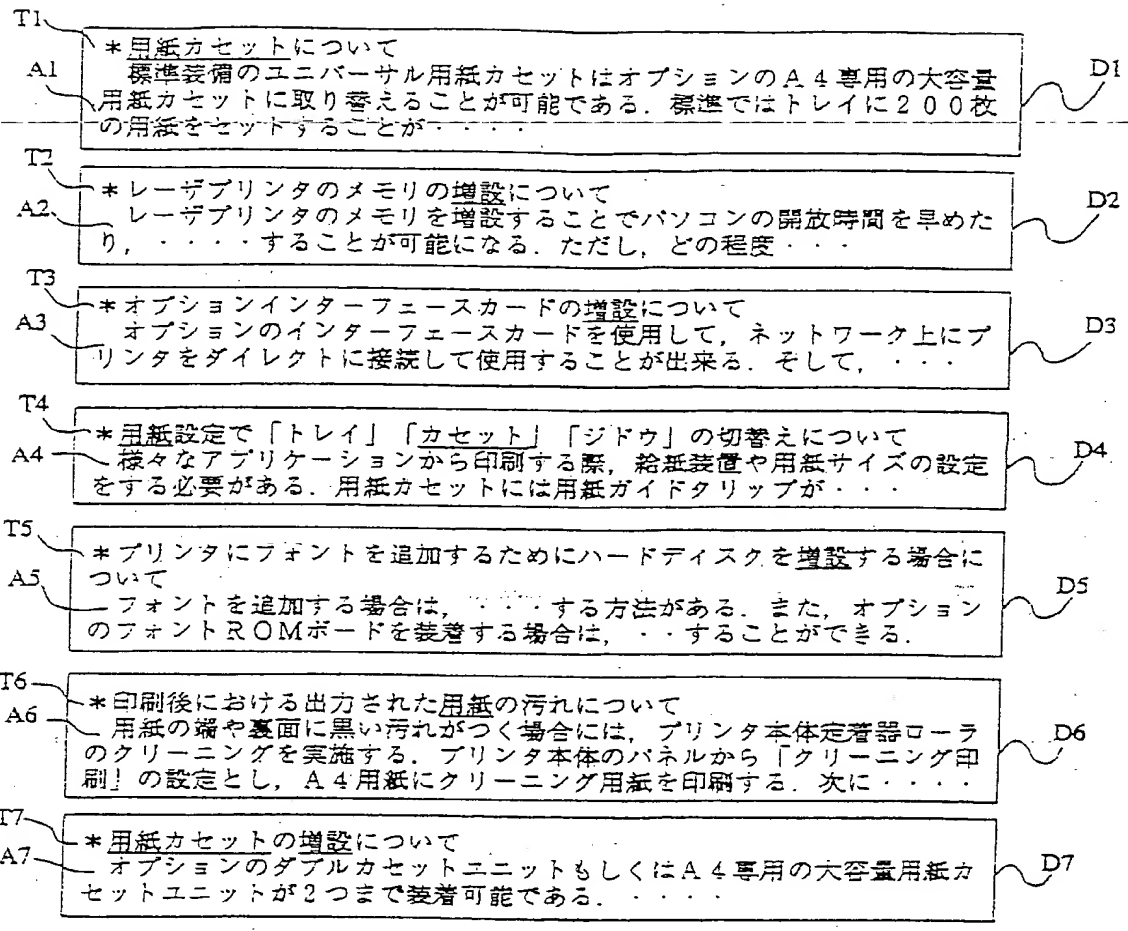
T1
＊用紙カセットについて
A1
標準装備のユニバーサル用紙カセットはオプションのＡ４専用の大容量用紙カセットに取り替えることが可能である．標準ではトレイに２００枚の用紙をセットすることが・・・・
D1
T2
＊レーザプリンタのメモリの増設について
A2
レーザプリンタのメモリを増設することでパソコンの開放時間を早めたり，・・・・することが可能になる．ただし，どの程度・・・
D2
T3
＊オプションインターフェースカードの増設について
A3
オプションのインターフェースカードを使用して，ネットワーク上にプリンタをダイレクトに接続して使用することが出来る．そして，・・・
D3
T4
＊用紙設定で「トレイ」「カセット」「ジドウ」の切替えについて
A4
様々なアプリケーションから印刷する際，給紙装置や用紙サイズの設定をする必要がある．用紙カセットには用紙ガイドクリップが・・・
D4
T5
＊プリンタにフォントを追加するためにハードディスクを増設する場合について
A5
フォントを追加する場合は，・・・する方法がある．また，オプションのフォントＲＯＭボードを装着する場合は，・・することができる．
D5
T6
＊印刷後における出力された用紙の汚れについて
A6
用紙の端や裏面に黒い汚れがつく場合には，プリンタ本体定着器ローラのクリーニングを実施する．プリンタ本体のパネルから「クリーニング印刷」の設定とし，Ａ４用紙にクリーニング用紙を印刷する．次に・・・・
D6
T7
＊用紙カセットの増設について
A7
オプションのダブルカセットユニットもしくはＡ４専用の大容量用紙カセットユニットが２つまで装着可能である．・・・・
D7

図 1 1

始め
文書群をクラスタに分類
1151
クラスタ間の関連性を判断
1152
関連性の高いクラスタを統合
1153
関連性の度合いを判定
1154
クラスタマージされる前の個々のクラスタに関する情報がわかるような表示内容の生成
1155
終り

図 1 2

特徴要素	文書D1	文書D2	文書D3	文書D4	文書D5	文書D6	文書D7
用紙	1			1		1	1
カセット	1			1			1
増設		1	1		1		1

図１３

特徴要素	クラスタ
用紙	Ｄ１，Ｄ４，Ｄ６，Ｄ７
カセット	Ｄ１，Ｄ４，Ｄ７
増設	Ｄ２，Ｄ３，Ｄ５，Ｄ７

図１４

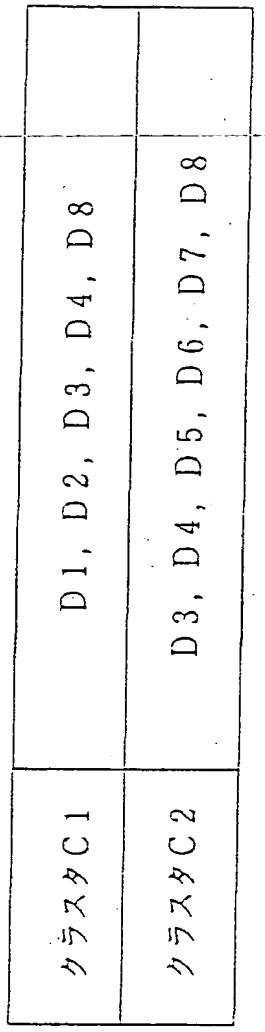

図１５

特徴要素	クラスタ	
用紙＋カセット	D1, D4, D6, D7	
増設	D2, D3, D5, D7	

図１６

クラスタ名	文書のタイトル
用紙カセット	＊用紙カセットについて ＊用紙設定で「トレイ」「カセット」「ジドウ」の切替えについて ＊印刷後における出力された用紙の汚れについて ＊用紙カセットの増設について

図１７

クラスタ名	文書のタイトル
用紙 カセット	＊用紙カセットについて ＊用紙設定で「トレイ」「カセット」「ジドウ」の切替えについて ＊印刷後における出力された用紙の汚れについて ＊用紙カセットの増設について

P.B.5818 – Patentlaan 2
2280 HV Rijswijk (ZH)
☎ +31 70 340 2040
TX 31651 epo nl
FAX +31 70 340 3016

Europäisches Patentamt
Zweigstelle in Den Haag Recherchen-abteilung

European Patent Office
Branch at The Hague Search division

Office européen des brevets
Département à La Haye Division de la recherche

Sturt, Clifford Mark
Miller Sturt Kenyon
9 John Street
London WC1N 2ES
GRANDE BRETAGNE

Datum/Date
09.06.04

Zeichen/Ref./Réf.	Anmeldung Nr./Application No./Demande n°./Patent Nr./Patent No./Brevet n°.
EPP13718A	00931690.2-2201-JP0003625

Anmelder/Applicant/Demandeur/Patentinhaber/Proprietor/Titulaire
SEIKO EPSON CORPORATION

COMMUNICATION

The European Patent Office herewith transmits as an enclosure the European search report for the above-mentioned European patent application.

If applicable, copies of the documents cited in the European search report are attached.

☐ Additional set(s) of copies of the documents cited in the European search report is (are) enclosed as well.

REFUND OF THE SEARCH FEE

If applicable under Article 10 Rules relating to fees, a separate communication from the Receiving Section on the refund of the search fee will be sent later.

European Patent Office

SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 00 93 1690

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	ZAMIR O ET AL: "WEB DOCUMENT CLUSTERING: A FEASIBILITY DEMONSTRATION" SIGIR '98. PROCEEDINGS OF THE 21ST ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. MELBOURNE, AUG. 24 - 28, 1998, ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RET, 1998, pages 46-54, XP002949043 ISBN: 1-58113-015-5	1-5, 10-13	G06F17/30
Y	* figure 1 * * page 48, left-hand column - page 49, right-hand column *	6-9	
Y	ZAMIR O ET AL: "GROUPER: A DYNAMIC CLUSTERING INTERFACE TO WEB SEARCH RESULTS" COMPUTER NETWORKS AND ISDN SYSTEMS, NORTH HOLLAND PUBLISHING. AMSTERDAM, NL, vol. 31, 17 May 1999 (1999-05-17), pages 1361-1374, XP002930746 ISSN: 0169-7552 * figure 2 * * page 1366, left-hand column * * page 1367, left-hand column - page 1368, left-hand column *	6-9	TECHNICAL FIELDS SEARCHED (Int.Cl.7) G06F
A	WILLETT P: "RECENT TRENDS IN HIERARCHIC DOCUMENT CLUSTERING: A CRITICAL REVIEW" INFORMATION PROCESSING & MANAGEMENT, ELSEVIER, BARKING, GB, vol. 24, no. 5, 1988, pages 577-597, XP000573921 ISSN: 0306-4573 * page 579 * * page 580 - page 582 *	1-13	

-/--

The supplementary search report has been based on the last set of claims valid and available at the start of the search.

Place of search	Date of completion of the search	Examiner
Munich	14 May 2004	Michalski, S

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C04)

European Patent Office

SUPPLEMENTARY EUROPEAN SEARCH REPORT

Application Number
EP 00 93 1690

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
A	CHANG C-H ET AL: "Customizable multi-engine search tool with clustering" COMPUTER NETWORKS AND ISDN SYSTEMS, NORTH HOLLAND PUBLISHING. AMSTERDAM, NL, vol. 29, no. 8-13, 1 September 1997 (1997-09-01), pages 1217-1224, XP004095318 ISSN: 0169-7552 * page 1220 - page 1221 *	1-13	
A	ZAMIR O ; ETZIONI O ; MADANI O ; KARP R M: "Fast and intuitive clustering of Web documents" PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 17 August 1997 (1997-08-17), pages 287-290, XP002280255 NEWPORT BEACH, CA, USA * page 289, left-hand column *	1,3,10, 12	
A	CROUCH C J ED - CHIARAMELLA Y ASSOCIATION FOR COMPUTING MACHINERY: "A CLUSTER-BASED APPROACH TO THESAURUS CONSTRUCTION" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL. (SIGIR). GRENOBLE, JUNE 13 - 15, 1988, NEW YORK, ACM, US, vol. CONF. 11, 13 June 1988 (1988-06-13), pages 309-320, XP000295046 * page 316 - page 317 *	1,3,10, 12	TECHNICAL FIELDS SEARCHED (Int.Cl.7)
A	SALTON G ET AL: "TERM-WEIGHTING APPROACHES IN AUTOMATIC TEXT RETRIEVAL" INFORMATION PROCESSING & MANAGEMENT, ELSEVIER, BARKING, GB, vol. 24, no. 5, 1988, pages 513-523, XP002035959 ISSN: 0306-4573 * page 516 - page 517 *	1,3,10, 12	

The supplementary search report has been based on the last set of claims valid and available at the start of the search.

Place of search	Date of completion of the search	Examiner
Munich	14 May 2004	Michalski, S

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C04)

WO 00/75810
PCT/JP00/03625

2 / 6
OP

PATENT COOPERATION TREATY

From the INTERNATIONAL BUREAU

PCT

NOTICE INFORMING THE APPLICANT OF THE COMMUNICATION OF THE INTERNATIONAL APPLICATION TO THE DESIGNATED OFFICES

(PCT Rule 47.1(c), first sentence)

To:
SUZUKI, Kisaburo
Intellectual Property Department
Seiko Epson Corporation
3-5, Owa 3-chome
Suwa-shi, Nagano 392-8502
JAPON

Date of mailing (day/month/year) 14 December 2000 (14.12.00)	
Applicant's or agent's file reference F005276WO00	**IMPORTANT NOTICE**

International application No.	**International filing date** (day/month/year)	**Priority date** (day/month/year)
PCT/JP00/03625	02 June 2000 (02.06.00)	04 June 1999 (04.06.99)

Applicant
SEIKO EPSON CORPORATION et al

1. Notice is hereby given that the International Bureau has communicated, as provided in Article 20, the international application to the following designated Offices on the date indicated above as the date of mailing of this Notice:
   AG,AU,DZ,KP,KR,US

   In accordance with Rule 47.1(c), third sentence, those Offices will accept the present Notice as conclusive evidence that the communication of the international application has duly taken place on the date of mailing indicated above and no copy of the international application is required to be furnished by the applicant to the designated Office(s).

2. The following designated Offices have waived the requirement for such a communication at this time:
   AE,AL,AM,AP,AT,AZ,BA,BB,BG,BR,BY,CA,CH,CN,CR,CU,CZ,DE,DK,DM,EA,EE,EP,ES,FI,GB,GD,
   GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,
   NO,NZ,OA,PL,PT,RO,RU,SD,SE,SG,SI,SK,SL,TJ,TM,TR,TT,TZ,UA,UG,UZ,VN,YU,ZA,ZW
   The communication will be made to those Offices only upon their request. Furthermore, those Offices do not require the applicant to furnish a copy of the international application (Rule 49.1(a-bis)).

3. Enclosed with this Notice is a copy of the international application as published by the International Bureau on
   14 December 2000 (14.12.00) under No. WO 00/75810

REMINDER REGARDING CHAPTER II (Article 31(2)(a) and Rule 54.2)

If the applicant wishes to postpone entry into the national phase until 30 months (or later in some Offices) from the priority date, **a demand for international preliminary examination** must be filed with the competent International Preliminary Examining Authority before the expiration of 19 months from the priority date.

It is the applicant's sole responsibility to monitor the 19-month time limit.

Note that only an applicant who is a national or resident of a PCT Contracting State which is bound by Chapter II has the right to file a demand for international preliminary examination.

REMINDER REGARDING ENTRY INTO THE NATIONAL PHASE (Article 22 or 39(1))

If the applicant wishes to proceed with the international application in the **national phase**, he must, within 20 months or 30 months, or later in some Offices, perform the acts referred to therein before each designated or elected Office.

For further important information on the time limits and acts to be performed for entering the national phase, see the Annex to Form PCT/IB/301 (Notification of Receipt of Record Copy) and Volume II of the PCT Applicant's Guide.

The International Bureau of WIPO
34, chemin des Colombettes
1211 Geneva 20, Switzerland

Facsimile No. (41-22) 740.14.35

Authorized officer
J. Zahra

Telephone No. (41-22) 338.83.38

PCT REQUEST

F005276WO00

0	**For receiving Office use only**	
0-1	International Application No.	
0-2	International Filing Date	
0-3	Name of receiving Office and "PCT International Application"	

0-4	**Form - PCT/RO/101 PCT Request**	
0-4-1	Prepared using	PCT-EASY Version 2.91 (updated 01.01.2001)
0-5	**Petition** The undersigned requests that the present international application be processed according to the Patent Cooperation Treaty	
0-6	**Receiving Office (specified by the applicant)**	Japanese Patent Office (RO/JP)
0-7	**Applicant's or agent's file reference**	F005276WO00
I	**Title of invention**	DOCUMENT CATEGORIZING METHOD, DOCUMENT CATEGORIZING APPARATUS, AND STORAGE MEDIUM ON WHICH A DOCUMENT CATEGORIZATION PROGRAM IS STORED
II	**Applicant**	
II-1	This person is:	applicant only
II-2	Applicant for	all designated States except US
II-4	Name	SEIKO EPSON CORPORATION
II-5	Address:	4-1, Nishi-Shinjuku 2-Chome Shinjuku-Ku, Tokyo 163-0811 Japan
II-6	State of nationality	JP
II-7	State of residence	JP
II-8	Telephone No.	03-3348-3114
II-9	Facsimile No.	03-3340-4258
III-1	**Applicant and/or inventor**	
III-1-1	This person is:	applicant and inventor
III-1-2	Applicant for	US only
III-1-4	Name (LAST, First)	NAGAISHI, Michihiro
III-1-5	Address:	c/o SEIKO EPSON CORPORATION 3-5, Owa 3-Chome Suwa-Shi, Nagano 392-8502 Japan
III-1-6	State of nationality	JP
III-1-7	State of residence	JP

PCT REQUEST

F005276WO00

III-2	**Applicant and/or inventor**	
III-2-1	This person is:	applicant and inventor
III-2-2	Applicant for	US only
III-2-4	Name (LAST, First)	MIWA, Shinji
III-2-5	Address:	c/o SEIKO EPSON CORPORATION 3-5, Owa 3-Chome Suwa-shi, Nagano 392-8502 Japan
III-2-6	State of nationality	JP
III-2-7	State of residence	JP
IV-1	**Agent or common representative; or address for correspondence** The person identified below is hereby/has been appointed to act on behalf of the applicant(s) before the competent International Authorities as:	agent
IV-1-1	Name (LAST, First)	SUZUKI, Kisaburo
IV-1-2	Address:	c/o Intellectual Property Department SEIKO EPSON CORPORATION 3-5, Owa 3-Chome Suwa-Shi, Nagano 392-8502 Japan
IV-1-3	Telephone No.	0266-52-3139
IV-1-4	Facsimile No.	0266-58-3243
IV-2	**Additional agent(s)**	additional agent(s) with same address as first named agent
IV-2-1	Name(s)	KAMIYANAGI, Masataka; SUZAWA, Osamu
V	**Designation of States**	
V-1	Regional Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	AP: GH GM KE LS MW SD SL SZ TZ UG ZW and any other State which is a Contracting State of the Harare Protocol and of the PCT (except MZ) EA: AM AZ BY KG KZ MD RU TJ TM and any other State which is a Contracting State of the Eurasian Patent Convention and of the PCT EP: AT BE CH&LI CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE and any other State which is a Contracting State of the European Patent Convention and of the PCT (except TR) OA: BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG and any other State which is a member State of OAPI and a Contracting State of the PCT

V-2	National Patent (other kinds of protection or treatment, if any, are specified between parentheses after the designation(s) concerned)	AE AG AL AM AT AU AZ BA BB BG BR BY CA CH&LI CN CR CU CZ DE DK DM DZ EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ UA UG US UZ VN YU ZA ZW
V-5	**Precautionary Designation Statement** In addition to the designations made under items V-1, V-2 and V-3, the applicant also makes under Rule 4.9(b) all designations which would be permitted under the PCT except any designation(s) of the State(s) indicated under item V-6 below. The applicant declares that those additional designations are subject to confirmation and that any designation which is not confirmed before the expiration of 15 months from the priority date is to be regarded as withdrawn by the applicant at the expiration of that time limit.	
V-6	**Exclusion(s) from precautionary designations**	NONE
VI-1	**Priority claim of earlier national application**	
VI-1-1	Filing date	04 June 1999 (04.06.1999)
VI-1-2	Number	11-158498(P)
VI-1-3	Country	JP
VI-2	**Priority claim of earlier national application**	
VI-2-1	Filing date	27 July 1999 (27.07.1999)
VI-2-2	Number	11-212501(P)
VI-2-3	Country	JP
VI-3	**Priority document request** The receiving Office is requested to prepare and transmit to the International Bureau a certified copy of the earlier application(s) identified above as item(s):	VI-1, VI-2
VII-1	**International Searching Authority Chosen**	Japanese Patent Office (JPO) (ISA/JP)

VIII	**Check list**	number of sheets	electronic file(s) attached
VIII-1	Request	4	–
VIII-2	Description	37	–
VIII-3	Claims	4	–
VIII-4	Abstract	1	f005276wo00.txt
VIII-5	Drawings	14	–
VIII-7	TOTAL	60	

PCT REQUEST

F005276WO00

	Accompanying items	paper document(s) attached	electronic file(s) attached
VIII-8	Fee calculation sheet	✓	-
VIII-9	Separate signed power of attorney	✓	-
VIII-16	PCT-EASY diskette	-	diskette
VIII-17	Other (specified):	Revenue stamps of transmittal fee for receiving office	-
VIII-18	**Figure of the drawings which should accompany the abstract**	1	
VIII-19	**Language of filing of the international application**	Japanese	
IX-1	**Signature of applicant or agent**		
IX-1-1	Name (LAST, First)	SUZUKI, Kisaburo	
IX-2	**Signature of applicant or agent**		
IX-2-1	Name (LAST, First)	KAMIYANAGI, Masataka	
IX-3	**Signature of applicant or agent**		
IX-3-1	Name (LAST, First)	SUZAWA, Osamu	

FOR RECEIVING OFFICE USE ONLY

10-1	**Date of actual receipt of the purported international application**	
10-2	**Drawings:**	
10-2-1	Received	
10-2-2	Not received	
10-3	**Corrected date of actual receipt due to later but timely received papers or drawings completing the purported international application**	
10-4	**Date of timely receipt of the required corrections under PCT Article 11(2)**	
10-5	**International Searching Authority**	ISA/JP
10-6	**Transmittal of search copy delayed until search fee is paid**	

FOR INTERNATIONAL BUREAU USE ONLY

11-1	**Date of receipt of the record copy by the International Bureau**	

EP・US　　特許協力条約

PCT

国際調査報告

（法８条、法施行規則第40、41条）
〔ＰＣＴ１８条、ＰＣＴ規則43、44〕

出願人又は代理人の書類記号　F005276W000	今後の手続きについては、国際調査報告の送付通知様式（ＰＣＴ／ＩＳＡ／２２０）及び下記５を参照すること。	
国際出願番号 ＰＣＴ／ＪＰ００／０３６２５	国際出願日（日.月.年）　０２．０６．００	優先日（日.月.年）　０４．０６．９９
出願人（氏名又は名称）　セイコーエプソン株式会社		

国際調査機関が作成したこの国際調査報告を法施行規則第41条（ＰＣＴ１８条）の規定に従い出願人に送付する。
この写しは国際事務局にも送付される。

この国際調査報告は、全部で　4　ページである。

☐ この調査報告に引用された先行技術文献の写しも添付されている。

１．国際調査報告の基礎

a．言語は、下記に示す場合を除くほか、この国際出願がされたものに基づき国際調査を行った。
☐ この国際調査機関に提出された国際出願の翻訳文に基づき国際調査を行った。

b．この国際出願は、ヌクレオチド又はアミノ酸配列を含んでおり、次の配列表に基づき国際調査を行った。
☐ この国際出願に含まれる書面による配列表
☐ この国際出願と共に提出されたフレキシブルディスクによる配列表
☐ 出願後に、この国際調査機関に提出された書面による配列表
☐ 出願後に、この国際調査機関に提出されたフレキシブルディスクによる配列表
☐ 出願後に提出した書面による配列表が出願時における国際出願の開示の範囲を超える事項を含まない旨の陳述書の提出があった。
☐ 書面による配列表に記載した配列とフレキシブルディスクによる配列表に記録した配列が同一である旨の陳述書の提出があった。

２．☐ 請求の範囲の一部の調査ができない（第Ⅰ欄参照）。

３．☒ 発明の単一性が欠如している（第Ⅱ欄参照）。

４．発明の名称は
☒ 出願人が提出したものを承認する。
☐ 次に示すように国際調査機関が作成した。

５．要約は
☒ 出願人が提出したものを承認する。
☐ 第Ⅲ欄に示されているように、法施行規則第47条（ＰＣＴ規則38.2(b)）の規定により国際調査機関が作成した。出願人は、この国際調査報告の発送の日から１カ月以内にこの国際調査機関に意見を提出することができる。

６．要約書とともに公表される図は、
第　1　図とする。☒ 出願人が示したとおりである。　☐ なし
☐ 出願人は図を示さなかった。
☐ 本図は発明の特徴を一層よく表している。

国際調査報告

国際出願番号 PCT／JP00／03625

第Ⅰ欄 請求の範囲の一部の調査ができないときの意見（第１ページの２の続き）

法第８条第３項（ＰＣＴ１７条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1. ☐ 請求の範囲 _____ は、この国際調査機関が調査をすることを要しない対象に係るものである。
つまり、

2. ☐ 請求の範囲 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、

3. ☐ 請求の範囲 _____ は、従属請求の範囲であってＰＣＴ規則6.4(a)の第２文及び第３文の規定に従って記載されていない。

第Ⅱ欄 発明の単一性が欠如しているときの意見（第１ページの３の続き）

次に述べるようにこの国際出願に二以上の発明があるとこの国際調査機関は認めた。

「複数の文書を意味的に共通性を有する複数のクラスタに分類したのちに、その複数のクラスタ間でそれぞれのクラスタに含まれる文書に基づいてそれらのクラスタの関連性を判断し、一定以上の関連性を有する少なくとも２つのクラスタを統合するクラスタマージ処理を行うこと」は、先行技術の域を出ないから、ＰＣＴ規則１３．２の第２文の意味において特別な技術事項でない。それ故、請求の範囲全てに共通の事項はない。
請求の範囲１－５，１０，１２は、クラスタの統合に関するものである。請求の範囲６－９，１１，１３は、新たなクラスタの表示に関するものである。

1. ☒ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求の範囲について作成した。

2. ☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求の範囲について調査することができたので、追加調査手数料の納付を求めなかった。

3. ☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったので、この国際調査報告は、手数料の納付のあった次の請求の範囲のみについて作成した。

4. ☐ 出願人が必要な追加調査手数料を期間内に納付しなかったので、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求の範囲について作成した。

追加調査手数料の異議の申立てに関する注意

☐ 追加調査手数料の納付と共に出願人から異議申立てがあった。

☒ 追加調査手数料の納付と共に出願人から異議申立てがなかった。

国際調査報告　　国際出願番号　PCT／JP00／03625

A.　発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

B.　調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷　G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1926－1996年
日本国公開実用新案公報	1971－2000年
日本国実用新案登録公報	1996－2000年
日本国登録実用新案公報	1994－2000年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）
ＪＩＣＳＴ科学技術文献ファイル　（（分類+クラスタ+cluster）＊（統合+merge））
ＷＰＩ　(cluster*merge)
ＩＮＳＰＥＣ　(cluster*merge)

C.　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X	ＪＰ，１０－１６２０１２，Ａ（松下電器産業株式会社） １９．６月．１９９８（１９．０６．９８） 全文	１－６， １０－１３
Y	全文 （ファミリーなし）	７－９

☒ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 17．08．00	国際調査報告の発送日 29.08.00
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員） 平井　誠　印　5L　9740 電話番号　03－3581－1101　内線　3560

様式PCT／ISA／210（第2ページ）（1998年7月）

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
Y	塩見、徳田、青山、柿ヶ原，「シソーラスを用いた文書データの自動分類法」，情報処理学会研究報告，Vol.97，No.4（97-NL-117）p.99-104，（日） ２０．１月．１９９７（２０．０１．９７） 特に、第100頁右欄下から13行～第101頁右欄4行	７－９
A	Oren Zamir, Oren Etzioni, "Grouper: a dynamic clustering interface to Web search results," Computer Networks, Vol.31, No.11-16, p.1361-1374 17.May.1999(17.05.99) 第1363頁右欄下から3行目～第1369頁右欄7行	１－１３